ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

THE VITALI COVERING THEOREM FOR CARATHÉODORY LINEAR MEASURE

BY JOHN F. RANDOLPH

(Received May 6, 1938)

Introduction

If A is an unrestricted plane set and F is a family of closed sets associated with A under certain conditions involving Lebesgue plane measure, the Vitali covering theorem states that there exists in F a sequence of its members which are mutually exclusive and whose union contains all of A except perhaps a subset of Lebesgue plane measure zero. In the statement of the Vitali covering theorem, if the conditions involving Lebesgue plane measure are replaced throughout by analogous conditions for Carathéodory linear measure, it is not known that the resulting statement is true for unrestricted sets A. In this paper it is proved, however, that this statement is true for all sets A (with finite measure) which satisfy one condition given in terms of Carathéodory linear measure. That this condition is not an unnatural restriction is shown by the fact that its analogue for Lebesgue measure is satisfied by all sets A.

In the ordinary Vitali covering theorem, the proof of sufficiency rests heavily on one particular key condition. Carathéodory's question as to whether this condition is necessary led H. Bohr ([4] p. 689) and Banach [1] to prove the answer to be in the affirmative. The theorem for linear measure keeps the analogous condition, the necessity of which is established in section 7.

For the reader's convenience the notations in Carathéodory's *Reelle Funktionen* [4] and in his Gött. Nach. article [5] have been followed. Discussion has been restricted to sets lying in the plane. Lebesgue measure will mean Lebesgue plane measure and linear measure will mean Carathéodory linear measure.

1. Statement of Theorem

The theorem we shall prove is stated as follows.

THEOREM 1. *Let A be a point set with L^*A positive and finite and U an arbitrary open set containing A. To each point P of A let there be associated a sequence of closed point sets*

$$\sigma_1(P), \quad \sigma_2(P), \cdots, \tag{1}$$

($\sigma_k(P)$ not necessarily including P) having the following characteristics:

a) *the point set $\sigma_k(P)$ lies in a circle $C_k(P)$ with center P whose radius $r_k(P)$ approaches zero with $1/k$,*

$$\lim_{k\to\infty} r_k(P) = 0; \tag{2}$$

b) *there exists a point set B with L^*B finite and a positive number $\alpha(P)$ depending on P, but independent of k such that*

$$L^*B\sigma_k(P) > \alpha(P)L^*AC_k(P); \tag{3}$$

If the inferior outer density of A is positive

$$\liminf_{\lambda\to 0} \frac{L^*Ac(P, r)}{2r} \equiv \underline{D}^*(A, P) > 0 \tag{†}$$

(where $c(P, r)$ is a circle with center P and radius r) for almost all of A, then there exists in A a sequence of points $P_1, P_2, \cdots$ and a corresponding sequence of integers $k_1, k_2, \cdots$ such that, with the notation,

$$\sigma_i' = \sigma_{k_i}(P_i), \qquad i = 1, 2, \cdots; \tag{4}$$

c) *the point sets σ_i' all lie within U and no two have a point in common*

$$\sigma_i' \subset U, \qquad \sigma_i'\sigma_j' = 0, \qquad i \neq j; \tag{5}$$

d) *and moreover the union*

$$s = \sigma_1' + \sigma_2' + \cdots \tag{6}$$

contains almost all of A

$$L^*AS = L^*A, \tag{7}$$

or

$$L^*(A - AS) = 0. \tag{8}$$

2. Remarks

Before proving this theorem we wish to show that if we replace linear measure throughout by Lebesgue measure we obtain a known theorem (theorem 1, [4] p. 299) with, to be sure, some redundant conditions. At first glance this statement may not appear to be true since in the theorem cited squares $w_k(P)$ are used instead of circles, the set B is not introduced, the inequality corresponding to (3) does not involve the set A, i.e. is written

$$m\sigma_k(P) > \alpha(P)mW_k(P) \tag{3'}$$

and finally the analogue for Lebesgue measure of the condition (†) on the set A is not given.

The option of using squares or circles is of course clear.

If now we require the existence of a set B such that

$$m^*B\sigma_k(P) > \alpha(P)mC_k(P) \tag{3''}$$

this condition would be unnecessary. For if this inequality held for some set B it would *a fortiori* hold if we replace B by the whole plane. It is, however, essential in the proof that $m\sigma_k(P)$ be finite, a property gained by requiring that $\sigma_k(P)$ be bounded, i.e. lie in a circle. We cannot, on the other hand, obtain

finiteness of $L\sigma_k(P)$ by merely requiring that $\sigma_k(P)$ be bounded. Thus one reason for introducing the set B is so the left member of (3) will be finite.

The set A does not appear on the right side of (3′), the inequality involving Lebesgue measure, as it does in the inequality (3) for linear measure. We may, however, require in place of (3′) the inequality

$$(3''') \qquad m\sigma_k(P) > \alpha(P)mAC_k(P)$$

which does involve A. For, when Lebesgue measure is used, the outer density of a set exists and is unity at almost all points of the set, i.e. the limit inferior and the limit superior of the ratio

$$\frac{m^*Ac(P, r)}{mc(P, r)}$$

are both unity at almost all points of A. Thus corresponding to P there is a number $K(P)$ such that

$$\frac{1}{2} < \frac{m^*AC_k(P)}{mC_k(P)} \leqq 1 \qquad \text{if } k > K(P)$$

for almost all points P of A. If we discard from the sequence $\sigma_k(P)$ all those members with $k \leqq K(P)$, then (3′) and (3‴) will be equivalent for the remaining sets of which there will be an infinite number associated with each point of A.

We have just used the fact that for Lebesgue measure, the inferior outer density of A is unity $\underline{D}^*(A, P) = 1$ at almost all points of A. This fact shows that the analogue of (†) for Lebesgue measure is automatically satisfied for all sets A.

3. Proof of a special case

We first prove a special case of theorem 1, namely

THEOREM 2. *In condition* (b), *α is taken independent of P as well as of k,*

$$(3_1) \qquad L^*B\sigma_k(P) > \alpha L^*AC_k(P),$$

and the outer density is taken bounded from zero

$$(\dagger_1) \qquad \underline{D}^*(A, P) \geqq \beta > 0$$

for almost all of A.

We may of course disregard all sets $\sigma_k(P)$ not lying entirely in U.

Let A_1 be the subset of A where the inferior outer density of A is less than β, i.e. $\underline{D}^*(A, P) < \beta$. Then $L^*A_1 = 0$ by hypothesis. As a permanent notation we shall let $B' = A + B$. Then L^*B' is finite and moreover from (3_1), $L^*B'\sigma_k(P) > \alpha L^*AC_k(P)$. Let A_2 be the subset of A where the superior outer density of B' is greater than unity, i.e. $D^*(B', P) > 1$. Then from a theorem proved by Sierpinski ([6], Theo. 1.) also $L^*A_2 = 0$. We may neglect the set $A_1 + A_2$ of linear measure zero and require that at each point of A, $\underline{D}^*(A, P) \geqq \beta$ and $\bar{D}^*(B', P) \leqq 1$.

We then retain for further consideration only those associated closed sets $\sigma_k(P)$ for which the corresponding circle $C_k(P)$ is such that

$$\frac{L^*AC_k(P)}{2r_k(P)} > \frac{\beta}{2}. \tag{9}$$

At each point of A this inequality holds for all sufficiently small radii so there remains an infinite number of sets $\sigma_k(P)$ associated with each point of A. In like manner we retain only those sets $\sigma_k(P)$ for which in addition

$$\frac{L^*B'C_k(P)}{2r_k(P)} < \frac{3}{2} \tag{10}$$

and still further only the remaining sets for which, with $q = (1 + 9/(2\alpha\beta))$,

$$\frac{L^*Ac(P, qr_k(P))}{2qr_k(P)} < \frac{3}{2}. \tag{11}$$

Even so, since $\bar{D}^*(B', P) \leqq 1$ and *a fortiori* $\bar{D}^*(A, P) \leqq 1$, we still have an infinite number of sets $\sigma_k(P)$ associated with each point of A. Furthermore for each of the retained sets $\sigma_k(P)$,

$$L^*B'\sigma_k(P) > 0, \tag{12}$$

since B' contains A and inequalities (9) and (3_1) hold.

Now (using the method employed by Banach [1]) let M be the least upper bound of $L^*B'\sigma_k(P)$ for all retained sets $\sigma_k(P)$. M is finite since L^*B' is finite. There is then a point P_1 of A and a number k_1 such that $\sigma_1' \equiv \sigma_{k_1}(P_1)$ is one of these retained sets and $L^*B'\sigma_1' > 2M/3$. Thus for any of the retained sets $\sigma_k(P)$

$$L^*B'\sigma_1' > (2/3)\, L^*B'\sigma_k(P).$$

We now discard all sets $\sigma_k(P)$ having a point in common with σ_1'. In general assuming the sets $\sigma_1', \cdots, \sigma_{n-1}'$ already defined we obtain a set $\sigma_n' \equiv \sigma_{k_n}(P_n)$ having no point in common with $s_{n-1} = \sum_{i=1}^{n-1} \sigma_i'$ and such that

$$L^*B'\sigma_n' > \tfrac{2}{3}\, L^*B\sigma_k(P)$$

for all sets $\sigma_k(P)$ having no point in common with s_{n-1}.

If this process of selection ends, say at the n^{th} stage, one will see that $A \subset S_n$.

If the process does not stop we notice first, since no two of the selected sets σ_i' have a point in common, that for each n, $L^*B' \geqq \sum_{i=1}^{n} L^*B'\sigma_i'$. This inequality with the inequalities (3_1) and (9) gives

$$L^*B' \geqq \alpha \sum_{i=1}^{n} L^*AC_{k_i}(P_i) > \alpha\beta \sum_{i=1}^{n} r_{k_i}(P_i).$$

Thus both of the series $\sum_{i=1}^{\infty} L^*B'\sigma_i'$ and $\sum_{i=}^{\infty} r_{k_i}(P_i)$ converge.

No set S_n includes the whole set A, but we assert that

$$A \subset S_{n-1} + \sum_{i=n}^{\infty} c(P_i, qr_{k_i}(P_i)) \text{ where } q = (1 + 9/(2\alpha\beta)). \tag{13}$$

For let P' be a point of A not in S_{n-1}. Then there is a number k' such that the closed set $\sigma_{k'}(P')$ has no point in common with S_{n-1}. Consequently a

number $j \geqq n$ exists such that $\sigma_{k'}(P')\sigma_j' \neq 0$. For otherwise we would have $L^*B'\sigma_j' > \frac{2}{3} L^*B'\sigma_{k'}(P') > 0$ for every j and this is impossible since the series $\sum L^*B'\sigma_j'$ has already been seen to converge. Let m be the smallest number j for which $\sigma_{k'}(P')\,\sigma_j' \neq 0$, i.e.

$$\sigma_{k'}(P')S_{m-1} = 0 \quad \text{while} \quad \sigma_{k'}(P')\sigma_m' \neq 0.$$

Thus $L^*B'\sigma_m' > (2/3)L^*B'\sigma_{k'}(P')$. Since $C_{k_m}(P_m) \supset \sigma_m'$ and the above inequality and (3_1) hold we have

$$L^*B'C_{k_m}(P_m) \geqq L^*B'\sigma_m' > \tfrac{2}{3} L^*B'\sigma_{k'}(P') > \tfrac{2}{3}\,\alpha L^*AC_{k'}(P').$$

But from (10) we have that the left member above is less than $3r_{k_m}(P_m)$ and from (9) the right member is greater than $(2/3)\alpha\beta r_{k'}(P')$. Thus

$$r_{k'}(P') < \frac{9}{2\alpha\beta}\, r_{k_m}(P_m).$$

But $\sigma_{k'}(P')\sigma_m' \neq 0$ so the distance from P' to P is less than or equal to

$$r_{k'}(P') + r_{k_m}(P_m) < \left(1 + \frac{9}{2\alpha\beta}\right) r_{k_m}(P_m) \equiv q r_{k_m}(P_m).$$

Consequently P' is a point of the circle $c(P_m\,,\ qr_{k_m}(P_m))$ and the assertion (13) is established.

Thus the part of A not included in S_{n-1} has outer linear measure less than or equal to

$$\sum_{i=n}^{\infty} L^*Ac(P_i,\, qr_{k_i}(P_i)).$$

This number, from (11), is in turn less than or equal to

$$3\sum_{i=n}^{\infty} qr_{k_i}(P_i) = 3q\sum_{i=n}^{\infty} r_{k_i}(P_i)$$

which converges to zero as n becomes infinite since the series $\sum r_{k_i}(P_i)$ converges.

This proves theorem 2. One will then see the

COROLLARY. *If all of the conditions of theorem 2 hold and ϵ is an arbitrary positive number, then there is an integer n such that*

$$L^*A(\sigma_1' + \cdots + \sigma_n') > L^*A - \epsilon.$$

4. Additional remarks

In the hypotheses of theorem 1 (as in the analogous theorem for Lebesgue measure) there is no specific requirement that any of the closed sets $\sigma_k(P)$ have even as much as one point in common with A. In this respect we know only that there is a set B having enough in common with $\sigma_k(P)$ to make (3) hold. The objection might be raised that no method is demonstrated for finding a suitable set B. Toward meeting this objection we first bring to light the hidden fact that $\sigma_k(P)$ and A do have points in common (at least for k

We then retain for further consideration only those associated closed sets $\sigma_k(P)$ for which the corresponding circle $C_k(P)$ is such that

$$\frac{L^*AC_k(P)}{2r_k(P)} > \frac{\beta}{2}. \tag{9}$$

At each point of A this inequality holds for all sufficiently small radii so there remains an infinite number of sets $\sigma_k(P)$ associated with each point of A. In like manner we retain only those sets $\sigma_k(P)$ for which in addition

$$\frac{L^*B'C_k(P)}{2r_k(P)} < \frac{3}{2} \tag{10}$$

and still further only the remaining sets for which, with $q = (1 + 9/(2\alpha\beta))$,

$$\frac{L^*Ac(P, qr_k(P))}{2qr_k(P)} < \frac{3}{2}. \tag{11}$$

Even so, since $\bar{D}^*(B', P) \leqq 1$ and *a fortiori* $\bar{D}^*(A, P) \leqq 1$, we still have an infinite number of sets $\sigma_k(P)$ associated with each point of A. Furthermore for each of the retained sets $\sigma_k(P)$,

$$L^*B'\sigma_k(P) > 0, \tag{12}$$

since B' contains A and inequalities (9) and (3_1) hold.

Now (using the method employed by Banach [1]) let M be the least upper bound of $L^*B'\sigma_k(P)$ for all retained sets $\sigma_k(P)$. M is finite since L^*B' is finite. There is then a point P_1 of A and a number k_1 such that $\sigma_1' \equiv \sigma_{k_1}(P_1)$ is one of these retained sets and $L^*B'\sigma_1' > 2M/3$. Thus for any of the retained sets $\sigma_k(P)$

$$L^*B'\sigma_1' > (2/3)\, L^*B'\sigma_k(P).$$

We now discard all sets $\sigma_k(P)$ having a point in common with σ_1'. In general assuming the sets $\sigma_1', \cdots, \sigma_{n-1}'$ already defined we obtain a set $\sigma_n' \equiv \sigma_{k_n}(P_n)$ having no point in common with $s_{n-1} = \sum_{i=1}^{n-1} \sigma_i'$ and such that

$$L^*B'\sigma_n' > \tfrac{2}{3}\, L^*B\sigma_k(P)$$

for all sets $\sigma_k(P)$ having no point in common with s_{n-1}.

If this process of selection ends, say at the n^{th} stage, one will see that $A \subset S_n$.

If the process does not stop we notice first, since no two of the selected sets σ_i' have a point in common, that for each n, $L^*B' \geqq \sum_{i=1}^{n} L^*B'\sigma_i'$. This inequality with the inequalities (3_1) and (9) gives

$$L^*B' \geqq \alpha \sum_{i=1}^{n} L^*AC_{k_i}(P_i) > \alpha\beta \sum_{i=1}^{n} r_{k_i}(P_i).$$

Thus both of the series $\sum_{i=1}^{\infty} L^*B'\sigma_i'$ and $\sum_{i=1}^{\infty} r_{k_i}(P_i)$ converge.

No set S_n includes the whole set A, but we assert that

$$A \subset S_{n-1} + \sum_{i=n}^{\infty} c(P_i, qr_{k_i}(P_i)) \text{ where } q = (1 + 9/(2\alpha\beta)). \tag{13}$$

For let P' be a point of A not in S_{n-1}. Then there is a number k' such that the closed set $\sigma_{k'}(P')$ has no point in common with S_{n-1}. Consequently a

number $j \geqq n$ exists such that $\sigma_{k'}(P')\sigma_j' \neq 0$. For otherwise we would have $L^*B'\sigma_j' > \frac{2}{3} L^*B'\sigma_{k'}(P') > 0$ for every j and this is impossible since the series $\sum L^*B'\sigma_j'$ has already been seen to converge. Let m be the smallest number j for which $\sigma_{k'}(P')\ \sigma_j' \neq 0$, i.e.

$$\sigma_{k'}(P')S_{m-1} = 0 \quad \text{while} \quad \sigma_k'(P')\sigma_m' \neq 0.$$

Thus $L^*B'\sigma_m' > (2/3)L^*B'\sigma_{k'}(P')$. Since $C_{k_m}(P_m) \supset \sigma_m'$ and the above inequality and (3_1) hold we have

$$L^*B'C_{k_m}(P_m) \geqq L^*B'\sigma_m' > \tfrac{2}{3} L^*B'\sigma_{k'}(P') > \tfrac{2}{3}\,\alpha L^*AC_{k'}(P').$$

But from (10) we have that the left member above is less than $3r_{k_m}(P_m)$ and from (9) the right member is greater than $(2/3)\alpha\beta r_{k'}(P')$. Thus

$$r_{k'}(P') < \frac{9}{2\alpha\beta}\, r_{k_m}(P_m).$$

But $\sigma_{k'}(P')\sigma_m' \neq 0$ so the distance from P' to P is less than or equal to

$$r_{k'}(P') + r_{k_m}(P_m) < \left(1 + \frac{9}{2\alpha\beta}\right) r_{k_m}(P_m) \equiv qr_{k_m}(P_m).$$

Consequently P' is a point of the circle $c(P_m\,,\ qr_{k_m}(P_m))$ and the assertion (13) is established.

Thus the part of A not included in S_{n-1} has outer linear measure less than or equal to

$$\sum_{i=n}^{\infty} L^*Ac(P_i,\, qr_{k_i}(P_i)).$$

This number, from (11), is in turn less than or equal to

$$3\sum_{i=n}^{\infty} qr_{k_i}(P_i) = 3q\sum_{i=n}^{\infty} r_{k_i}(P_i)$$

which converges to zero as n becomes infinite since the series $\sum r_{k_i}(P_i)$ converges.

This proves theorem 2. One will then see the

COROLLARY. *If all of the conditions of theorem 2 hold and ϵ is an arbitrary positive number, then there is an integer n such that*

$$L^*A(\sigma_1' + \cdots + \sigma_n') > L^*A - \epsilon.$$

4. Additional remarks

In the hypotheses of theorem 1 (as in the analogous theorem for Lebesgue measure) there is no specific requirement that any of the closed sets $\sigma_k(P)$ have even as much as one point in common with A. In this respect we know only that there is a set B having enough in common with $\sigma_k(P)$ to make (3) hold. The objection might be raised that no method is demonstrated for finding a suitable set B. Toward meeting this objection we first bring to light the hidden fact that $\sigma_k(P)$ and A do have points in common (at least for k

large enough) and then show that with no loss of generality whatever the set B may be discarded entirely if requirement b) is replaced by a similar condition involving A only.

First then

LEMMA. *Under the hypotheses of theorem* 2, *there is a set* $A_1 \subset A$ *with* $L^*A_1 = L^*A$ *and for each point* P *of* A' *a number* $K(P)$ *such that*

$$L^*A\sigma_k(P) > \alpha\left(1 - \frac{\beta}{2}\right) L^*AC_k(P) \qquad \text{if } k \geqq K(P).$$

Let $\bar{A}$ be the *massgleiche Hülle* of A and $\bar{B}'$ of B'. Since $\bar{A}$ and $\bar{B}'$ are both measurable and contain A we may take $\bar{A} \subset \bar{B}'$. Then $L\bar{B}'\sigma_k(P) = L\bar{A}\sigma_k(P) + L(\bar{B}' - \bar{A})\sigma_k(P)$. But from the fact that (3_1) holds, that ([4] theorem 12, p. 269) $L\bar{A}\sigma_k(P) = L^*A\sigma_k(P)$, and that $\sigma_k(P) \subset C_k(P)$ we then have

$$\alpha L^*AC_k(P) < L^*A\sigma_k(P) + L(\bar{B}' - \bar{A})C_k(P). \tag{14}$$

But since the set $\bar{B}' - \bar{A}$ is linearly measurable with finite linear measure, its superior density is zero at almost all points of its complement (see [2] theorem 2 or [5] theorem 4) so, in particular $\bar{D}(\bar{B}' - \bar{A}, P) = 0$ for all points P of a set A_2 contained in A and having $L^*A_2 = L^*A$. Using in addition condition $(\dagger_1)$ of theorem 2, we have that there exists a subset A_1 of A with $L^*A_1 = L^*A$ such that to each point P of A_1 there is a number $K(P)$ so

$$\frac{L(\bar{B}' - \bar{A})C_k(P)}{2r_k(P)} < \frac{\alpha\beta^2}{4} \quad \text{and} \quad \frac{L^*AC_k(P)}{2r_k(P)} > \frac{\beta}{2}$$

hold simultaneously if P belongs to A_1 and $k \geqq K(P)$. This last statement with (14) gives the desired result.

Next the existence of the set B may be discarded entirely. For

THEOREM 3. *A theorem equivalent to theorem* 2 *is obtained if condition* b) *is replaced by*

$\bar{\text{b}}$) *there exists a positive number* $\bar{\alpha}$ *independent of* P *and* k *such that*

$$L^*A\sigma_k(P) > \bar{\alpha}L^*AC_k(P). \tag{$\bar{3}$}$$

For if $\bar{\text{b}}$) holds we know the existence of a suitable set B of b); namely A itself. Conversely if b) holds then condition $\bar{\text{b}}$) follows from the above lemma (after all sets $\sigma_k(P)$ with $k \leqq K(P)$ are discarded) with $\bar{\alpha} = \alpha\left(1 - \frac{\beta}{2}\right)$.

5. Proof of theorem 1

First, the condition on A that the inferior outer density be bounded from zero by a positive constant may be replaced by the weaker condition of theorem 1 that this same density function be merely positive for almost all of A.

For let A_n be the part of A where $\underline{D}^*(A, P) \geqq \frac{1}{n}$. Then $A_1 \subset A_2 \subset \cdots$

and $V = A_1 + A_2 + \cdots$ where $L^*V = L^*A$ so $L^*A = \lim_{n\to\infty} L^*A_n$. Consequently corresponding to an ϵ there is a number m such that

$$L^*A_m > L^*A - \epsilon/2. \tag{15}$$

It has been proved ([5] theorem 9) that if G is a set with L^*G finite and $\underline{G}$ is any subset of G, then G and $\underline{G}$ have the same inferior outer densities at almost all points of $\underline{G}$. Consequently at almost all points of A_m its inferior outer density is the same as that of A so is bounded fron zero by the positive constant $1/m$. The set A_m then satisfies the conditions of theorem 2 so from the corollary there is, among the sets $\sigma_k(P)$ associated with points of A_m, a finite number of mutually exclusive sets $\sigma_{k_1}(P_1), \cdots, \sigma_{k_n}(P_n)$ whose unions s_1 lies in the given neighborhood U of A and is such that $L^*A_m s_1 > L^*A_m - \epsilon/2$. Since $A \supset A_m$, $L^*As_1 > L^*A_m - \epsilon/2$ which with (15) gives $L^*As_1 > L^*A - \epsilon$ or since s_1 is linearly measurable

$$L^*(A - As_1) < \epsilon_1. \tag{16}$$

Since s_1 is the union of a finite number of closed sets it is itself closed and since it lies in U (the open set containing A), $U - s_1$ is open. Let $A' = A - As_1$ and $U^1 = U - s_1$. Then $\underline{D}^*(A', P) = \underline{D}^*(A, P)$ at all points of A' and thus is positive at almost all points of A'. One will see that in the same manner we used to determine the set s_1 in relation to A we may now determine s_2 in relation to A', where s_2 is the union of a finite number of mutually exclusive sets $\sigma_k(P)$ lying in U' and such that $L^*(A' - A's_2) < \frac{\epsilon}{2}$. But $A' - A's_2 = A - A(s_1 + s_2)$ so we have

$$L^*[A - A(s_1 + s_2)] < \frac{\epsilon}{2}. \tag{17}$$

We see that we may determine in succession a sequence of mutually exclusive sets $s_1, s_2, \cdots$ each the union of a finite number of mutually exclusive sets $\sigma_k(P)$ that lie in U such that

$$L^*[A - A(s_1 + s_2 + \cdots)] \leqq L^*[A - A(s_1 + \cdots + s_n)] < \epsilon/n$$

for each n. Thus the inferior outer density may be merely positive instead of actually bounded from zero.

Next we see that the positive constant α of theorem 2 may be replaced by a function positive at almost all points P of A. For if we let A_n be the part of A where $\alpha(P) > 1/n$ the above argument may be repeated verbatim and lead to the stated conclusion.

6. Additivity and complete additivity

THEOREM 4. *If A_1 and A_2 are two sets each satisfying the conditions of theorem* 1, *then $A_1 + A_2$ also satisfy these conditions.*

The only trouble comes in showing from the inequalities

$$L^*A_i\sigma_k(P) > \alpha_i(P)L^*A_iC_k(P),\ \alpha_i(P) > 0, \qquad i = 1, 2, \tag{18}$$

that for almost all points P of $A_1 + A_2$ there is a number $\alpha(P) > 0$ such that

$$L^*(A_1 + A_2)\sigma_k(P) > \alpha(P)L^*(A_1 + A_2)C_k(P).$$

At a point P let $\beta_i(P)$, $i = 1, 2$, be the inferior outer densities of A_1 and A_2 respectively. Then at almost all points P of $A_1 + A_2$ either $\alpha_1(P)$ or $\alpha_2(P)$ is not zero. Also at almost all P of A_i there is a number $K(P)$ such that if $k > K(P)$

$$\frac{L^*A_iC_k(P)}{2r_k(P)} > \frac{\beta_i(P)}{2}, \quad i = 1, 2 \quad \text{and} \quad \frac{L^*(A_1 + A_2)C_k(P)}{2r_k(P)} < \frac{3}{2}. \tag{19}$$

From (18) and the first inequality of (19), for almost all P of A_i

$$LA_i\sigma_k(P) > \alpha_i(P)\beta_i(P)r_k(P), \quad i = 1, 2 \text{ if } k > K(P),$$

With $3\alpha(P) = \max_{i=1,2}[\alpha_i(P)\cdot\beta_i(P)]$ we have that $\alpha(P) > 0$ at almost all points of $A_1 + A_2$ and using the second inequality of (19)

$$L^*(A_1 + A_2)\sigma_k(P) > 3\,\alpha(P)r_k(P) > \alpha(P)L^*(A_1 + A_2)C_k(P)$$

for almost of P of $A_1 + A_2$ if $k \geqq K(P)$ which is the desired result.

We now have, for a set A which may have infinite measure,

THEOREM 5. *If A is the union of sets $A_1, A_2, \cdots$ each satisfying the conditions of theorem* 1 *and H is an open set containing A, then there exists among the associated closed sets a sequence of mutually exclusive sets whose union s lies in H and $L(A - As) = 0$.*

Among the closed sets associated with A_1 there is a finite number whose union s_1 lies in H and $L^*(A_1 - A_1s_1) < 1$. Then each of the sets $A_1 - A_1s_1$ and $A_2 - A_2s_1$ with the closed sets associated with their points again satisfies theorem 1 and thus (from the additivity expressed in theorem 4) their union $A_2' = (A_1 + A_2) - (A_1 + A_2)s_1$ again satisfies the conditions of theorem 1. Thus among the closed sets associated with points of A_2' there exists a finite number of mutually exclusive sets whose union s_2 lies in the open set $H - s_1$ and $L^*(A_2' - A_2's_2) < \frac{1}{2}$, i.e.

$$L^*[(A_1 + A_2) - (A_1 + A_2)\cdot(s_1 + s_2)] < \tfrac{1}{2}.$$

Thus there exists a sequence of mutually exclusive sets $s_1, s_2, \cdots$ each the union of a finite number of closed sets such that for each n

$$L^*[(A_1 + \cdots + A_n) - (A_1 + \cdots + A_n)\cdot(s_1 + \cdots + s_n)] < 1/n.$$

With $s = s_1 + s_2 + \cdots$, then $(A_1 + \cdots + A_n) - (A_1 + \cdots + A_n)\cdot s$ is contained in the above set and increases with n to $A - As$ so $L^*(A - As) = 0$.

7. Necessity

Analogously to Carathéodory's question mentioned in the introduction, one might ask whether the conditions on the closed sets $\sigma_k(P)$ could be replaced by the condition: Each set $\sigma_k(P)$ is a rectangle with center P and longest side decreasing to zero with $1/k$. The answer is in the negative as the following example shows.

Let c_1 be a unit circle. Divide the circumference of c_1 into 4 equal parts and tangent internally to this circumference at the points of division draw a circle of radius $1/2^2$. Call these circles $c_{2,1}, \cdots, c_{2,4}$ and D_2 their union. At the n^{th} stage draw $m_n = 2^2 \cdot 3^2 \cdot \cdots \cdot n^2$ circles $c_{n,1}, \cdots, c_{n,m_n}$ by dividing the circumference of each circle $c_{n-1,j}$ into n^2 equal parts and at each point if division draw tangent internally a circle $c_{n,i}$ of radius

$$\frac{1}{n^2}\left[\frac{1}{2^2, \cdots (n-1)^2}\right] = \frac{1}{(n!)^2}.$$

Let D_n be the union of these m_n circles.

The set A is the intersection $A = D_2 \cdot D_3 \cdot \cdots$. The exact measure of A is immaterial, but we must know (as shown by Besicovitch [2]) that $0 < LA < \infty$. By symmetry $0 < LAc_{n,i} = LAc_{n,j} = LA/(n!)^2 < \infty$.

Let P be a point of A. For every k this point P belongs to some circle $c_{k,i}$ which in turn lies in some $c_{k-1,j}$. Let $\sigma_k(P)$ be the smallest rectangle that has its center at P, includes all of $c_{k,i}$ and the center of $c_{k-1,j}$ and has long sides parallel to the line joining the centers of the circles $c_{k,i}$ and $c_{k-1,j}$. Thus the width of $\sigma_k(P)$ is less than or equal to $4/(k!)$ (i.e. twice the diameter of $c_{k,i}$) and length less than or equal to $2/[(k-1)!]^2$. Also the distance between the line of centers of $c_{k,i}$ and $c_{k-1,j}$ and either of the sides of $\sigma_k(P)$ parallel to it is $\leqq 3/(k!)^2$.

Let $c_{k,p}$ and $c_{k,q}$ be two neighboring circles of a circle $c_{k-1,j}$. The distance from $c_{k,p}$ to the line joining the centers of $c_{k,q}$ and $c_{k-1,j}$ is then

$$\left[\frac{1}{[(k-1)!]^2} - \frac{1}{(k!)^2}\right] \sin\frac{2\pi}{k^2} \geqq \frac{1}{[(k-1)!]^2}\left[1 - \frac{1}{k^2}\right]\frac{2}{\pi}\cdot\frac{2\pi}{k^2}$$

$$= \frac{4}{(k!)^2}\left[1 - \frac{1}{k^2}\right] > \frac{3}{(k!)^2} \quad \text{if} \quad k > 2.$$

Thus if $k \geqq 3$, $\sigma_k(P)$ contains all of one circle $c_{k,i}$ and no point of any other circle $c_{k,j}$, $j \neq i$.

Now let $\sigma_{k_1}(P_1)$, $\sigma_{k_2}(P_2)$, $\cdots$ be a sequence with every $k_i \geqq 3$ of non-overlapping rectangles selected from the family of rectangles associated with the set A in the above manner. Furthermore let $k_1 \leqq k_2 \leqq \cdots$. Then because no two of these rectangles have a point in common, we have that the first 2^2 of the k_i's is each $\geqq 3$, the next $2^2 \cdot 3^2$ of the k's is each $\geqq 4$, etc., i.e.

$$LA \sum_{i=1}^{4} \sigma_{k_i}(P_i) \leqq \frac{2^2 LA}{2^2 \cdot 3^2}, \qquad LA \sum_{i=5}^{5+2^2\cdot 3^2} \sigma_{k_i}(P_i) \leqq \frac{(3!)^2 LA}{(4!)^2},$$

etc. Thus

$$L\left[A\sum_{i=1}^{\infty}\sigma_{k_i}(P_i)\right] \leqq \left(\sum_{n=3}^{\infty}\frac{1}{n^2}\right)LA,$$

so

$$L\left[A - A\sum_{i=1}^{\infty}\sigma_{k_i}(P_i)\right] \geqq \left(1 - \sum_{n=3}^{\infty}\frac{1}{n^2}\right)LA > 0,$$

i.e. no mutually exclusive sequence of these rectangles can include almost all of A.

INSTITUTE FOR ADVANCED STUDY.

BIBLIOGRAPHY

(1) BANACH, S.: *Sur le Théorème de M. Vitali.* Funda. Math. vol. 5 (1924) p. 130.
(2) BESICOVITCH, A. S.: *On the fundamental geometrical properties of linearly measurable plane sets of points.* Math. Ann. vol. 98 (1927) p. 422.
(3) CARATHÉODORY, C.: *Über das lineare Mass von Punktmengen.* Gött. Nach. (1914) p. 404.
(4) CARATHÉODORY, C.: *Reelle Funktionen.* 1927.
(5) RANDOLPH, J. F.: *Some Density Properties of Point Sets.* Ann. of Math. vol. 37 (1936) p. 336.
(6) SIERPIŃSKI, W.: *Sur la densité linéaire des ensembles plans.* Funda. Math. vol. 9 (1927) p. 172.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

METRIC LATTICES*

BY L. R. WILCOX AND M. F. SMILEY

(Received June 1, 1938)

Introduction

It is well known that a linearly ordered set L possesses a topology defined purely in terms of the ordering relation $<$ in the following manner. Let $(a_i; i = 1, 2, \cdots)$ be a sequence in L; then $\lim^* a_i = a$ is defined to mean the existence of sequences $(a_i'; i = 1, 2, \cdots)$, $(a_i''; i = 1, 2, \cdots)$, where $a_i' \leqq a_j'$, $a_i'' \geqq a_j''$ for $i < j$, such that $a_i' \leqq a_i \leqq a_j''$ for $i = 1, 2, \cdots$, and such that[1]

$$\text{l.u.b. } [a_i'; i = 1, 2, \cdots] = a = \text{g.l.b. } [a_i''; i = 1, 2, \cdots].$$

A natural definition of *completeness* of L in this topology is that for every set $S \subset L$ both l.u.b. S and g.l.b. S should exist. A smallest complete extension $\bar{L}$ of L exists and can be constructed by the usual method employing "Dedekind cuts." If L is an interval of rational numbers, then it is well known that $\bar{L}$ is the real interval with the same endpoints. In this example, however, L is also a metric space (with metric $| a - b |$), and its *metric-completion* (in the sense of Cantor[2]) is also $\bar{L}$.

This discussion leads us to the problem of this paper, viz., to find general sufficient conditions under which a partially ordered set which is at the same time a metric space has an order completion and a metric completion which are isomorphic. We shall consider only the case where L is a lattice, and shall assume throughout certain natural relations between the metric and the ordering relation. Under these assumptions we shall determine necessary and sufficient conditions for the equivalence of the two completions of L. Our metric will be specialized to a considerable extent; indeed, we prefer to introduce first a "norm" and then show how a metric may be derived from this norm in a manner resembling the introduction of the "dimensional metric" in continuous geometries.[3] Many of our results, particularly our lemmas, are restatements or

---

* Presented to the American Mathematical Society December 28, 1937.

[1] It is readily verified that for real numbers this topology coincides with the usual topology, since it amounts to defining convergence as equality of upper and lower limits. For a generalization to continuous geometries, cf. J.v. Neumann, *Examples of continuous geometries*, Proc. Nat. Acad. Sci., 22 (1936), p. 107.

[2] F. Hausdorff, *Mengenlehre*, 3rd ed., Berlin and Leipzig, 1935, pp. 106–107.

[3] J. v. Neumann, op. cit., p. 106. Cf. also V. Glivenko, *Contributions à l'étude des systèmes de choses normées*, Amer. Journal of Math., 59 (1937), pp. 941–956; and *Géométrie des systèmes de choses normées*, Amer. Journal of Math., 58 (1936), pp. 799–828. In the last two papers investigation is made of systems which are somewhat more special than those discussed in this paper.

generalizations of results contained in the theory of continuous geometries, but inasmuch as the details of this theory are as yet not accessible, we shall give the proofs in full without explicit reference to the work of von Neumann.

We shall employ the usual set-theoretical notations, with the following exceptions. For a function $f(x)$ we denote the *system* of all elements $f(x)$ for which x has the property $\epsilon(x)$ by $(f(x);\ \epsilon(x))$. The *set* of these elements is $[f(x);\ \epsilon(x)]$. If S is a class of sets, then the set-theoretical sum (intersection) of the sets in S is denoted by $\mathfrak{S}\ S$ ($\mathfrak{P}\ S$). The empty set is denoted by Θ. If L is a set of elements $a, b, c, \cdots$, and $<$ is a binary relation defined for all pairs (a, b) with $a, b, \epsilon L$, we say that L is a *partially ordered set* (with respect to $<$) in case (1) never $a < a$, (2) $a \leqq b, b \leqq c$ implies $a \leqq c$, and (3) $a \leqq b, b \leqq a$ implies $a = b$; L is an *additive* (*multiplicative*) *system* in case every two elements $a, b \epsilon L$ possess a *least upper bound* $a + b \epsilon L$ (*greatest lower bound* $ab \epsilon L$) characterized by $x \geqq a + b$ if and only if $x \geqq a, b$ ($x \leqq ab$ if and only if $x \leqq a, b$); L is a *lattice* in case it is both an additive and a multiplicative system. If L is a partially ordered set, then a *zero* 0 (*unit* 1) is an element of L characterized by $0 \leqq a$ ($1 \geqq a$) for every $a \epsilon L$. These elements are unique if existent; their existence is assumed only where explicit statement is made to this effect. A lattice is *continuous* if every set $S \subset L$ possesses a *sum* (least upper bound) $\sum S$ such that $x \geqq \sum S$ if and only if $x \geqq y$ for each $y \epsilon S$, and a *product* (greatest lower bound) $\prod S$ such that $x \leqq \prod S$ if and only if $x \leqq y$ for each $y \epsilon S$; $\sum S$ and $\prod S$ are unique if existent. If L is a partially ordered set, we say that a sequence $(a_i;\ i = 1, 2, \cdots)$ is *increasing* (*decreasing*) in case $a_i \leqq a_j (a_i \geqq a_j)$ for $i < j$. Familiarity with the elements of the theory of lattices on the part of the reader is presupposed.[4]

Preparations for the theory are given in Section 1 where a *norm* is introduced and its properties studied. Relations between the two types of completeness of the systems introduced in Section 1 are examined in Section 2. The interdependence of the two completions of these systems is treated in Section 3. In particular, we find in this section, under mild hypotheses, necessary and sufficient conditions that the two completions be isomorphic. Finally, a general example of our abstract systems is discussed in Section 4.

1. Normed Lattices

In this section we shall study the relations between a real valued function $D(a)$ defined on an additive system L and a function $\delta_D(a, b)$ defined in terms of $D(a)$ by (1.1) (cf. Definition 1.1). A characterization is given of the functions $D(a)$ having the property that $\delta_D(a, b)$ is a metric. Finally, we find, under mild hypotheses, necessary and sufficient conditions on a metric $\delta(a, b)$ for L that there exist a function $D(a)$ such that $\delta = \delta_D$.

[4] Cf. for example G. Birkhoff, *On the combination of subalgebras*, Proc. Cambridge Phil. Soc., 29 (1933), pp. 441–464; O. Ore, *On the foundation of abstract algebra* I, Annals of Math., 36 (1935), pp. 406–437.

DEFINITION 1.1: (a) Let L be an additive system, and let $D(a)$ be a real-valued function defined on L. Then we define

$$\delta_D(a, b) \equiv 2D(a + b) - D(a) - D(b). \tag{1.1}$$

We say that $D(a)$ is *monotone increasing* (*decreasing*) in case $D(a) \leqq D(b)$ ($D(a) \geqq D(b)$) for $a < b$; $D(a)$ is *properly monotone increasing* (*decreasing*) in case $D(a) < D(b)$ ($D(a) > D(b)$) for $a < b$. The function $D(a)$ is a *norm* in case $\delta_D(a, b)$ is a metric for L; in this case we say that L is *normed* (by $D(a)$).

(b) If L is a lattice, then a real-valued function $D(a)$, $a \in L$ is an *affine dimension function* in case $D(a)$ is properly monotone increasing and

$$D(a + b) + D(ab) \leqq D(a) + D(b); \tag{1.2}$$

$D(a)$ is a *dimension function* in case $D(a)$ is properly monotone increasing and

$$D(a + b) + D(ab) = D(a) + D(b).$$

(c) Let L be an additive system, and let $\delta(a, b)$ be real-valued and defined for every $a, b \in L$; then we define

$$\Delta(a, b, c) \equiv \tfrac{1}{2}[\delta(a, b) + \delta(b, c) - \delta(a, c)].$$

When $\delta = \delta_D$ for a given $D(a)$, we write $\Delta_D(a, b, c)$.

LEMMA 1.1: *If $D(a)$ is a real-valued function defined on an additive system L, and if $\delta(a, b) \equiv \delta_D(a, b)$, then*

(1) $D(a) - D(b) = \delta(a, b)$ *if* $a \geqq b$;

(2) *if $D(a)$ is monotone increasing, then* $| D(a) - D(b) | \leqq \delta(a, b)$;

(3) $\delta(a, b) = \delta(b, a)$, $\delta(a, a) = 0$;

(4) $\Delta(a, a + b, b) = 0$;

(5) *$D(a)$ is monotone increasing if and only if* $\delta(a, b) \geqq 0$;

(6) *$D(a)$ is properly monotone increasing if and only if* $\delta(a, b) > 0$ *for* $a \neq b$; *in this case* $a \leqq b$ *with* $D(a) = D(b)$ *implies* $a = b$.

PROOF: Parts (1) and (3) are obvious. Clearly $2D(a + b) \geqq 2D(a)$, $2D(a + b) \geqq 2D(b)$ if $D(a)$ is monotone increasing. Thus

$$\delta(a, b) - (D(a) - D(b)) = 2D(a + b) - 2D(a) \geqq 0,$$

$$\delta(a, b) - (D(b) - D(a)) = 2D(a + b) - 2D(b) \geqq 0.$$

This proves (2). Part (4) follows from (1) since $\delta(a, a + b) + \delta(a + b, b) = D(a + b) - D(a) + D(a + b) - D(b) = \delta(a, b)$. The converse implications in (5) and (6) follow from (1). The forward implication in (5) follows from (2). To prove the first of the forward implications in (6) let $a \neq b$. Clearly either $a + b > a$ or $a + b > b$, and $D(a + b) > D(a)$ or $D(a + b) > D(b)$. Thus in either case $\delta(a, b) = D(a + b) - D(a) + D(a + b) - D(b) > 0$. The final implication in (6) is obvious. The proof is now complete.

LEMMA 1.2: *If $D(a)$ is a real-valued function defined on an additive system L, and if $\delta(a, b) \equiv \delta_D(a, b)$, then*

(1) $\Delta(a, b, c) = D(a + b) + D(b + c) - D(a + c) - D(b)$;
(2) $\Delta(a, b, c) = \delta(a, b) - \delta(a + c, b + c)$ *for* $b \leqq a$;
(3) $\delta(a, b) - \delta(a + c, b + c) = \Delta(a + b, a, c) + \Delta(a + b, b, c)$;
(4) $\Delta(a, b, c) \geqq 0$ *for every* $a, b, c \in L$ *implies the following equivalent statements*:
(i) $\delta(a + c, b + c) \leqq \delta(a, b)$ *for every* $a, b, c \in L$,
(ii) $\delta(a + c, b + c) \leqq \delta(a, b)$ *for* $b \leqq a$,
(iii) $D(a + c) + D(b) \leqq D(a) + D(c + b)$ *for* $b \leqq a$,
(iv) $\delta(a + c, b + d) \leqq \delta(a, b) + \delta(c, d)$;
(5) *if* $D(a)$ *is monotone increasing, then* $\Delta(a, b, c) \geqq 0$ *if and only if one of the equivalent statements* (i)–(iv) *holds.*

PROOF: Parts (1) and (3) follow from Definition 1.1 by an easy computation. Part (2) follows from (1). Part (4) is now evident except for the fact that (ii) implies (i). Using Lemma 1.1 (4) we obtain

$$\delta(a, b) - \delta(a + c, b + c) = \delta(a, a + b) + \delta(a + b, b) - (\delta(a + c, a + b + c) + \delta(a + b + c, b + c)) = (\delta(a, a + b) - \delta(a + c, a + b + c)) + (\delta(a + b, b) - \delta(a + b + c, b + c)) \geqq 0.$$

To prove (5) we show that (iii) implies that $\Delta(a, b, c) \geqq 0$ for every $a, b, c \in L$. By (1) and (4) (iii) we have

$$\Delta(a, b, c) = D(a + b) + D(b + c) - D(a + c) - D(b) \geqq D(a + b) - D(b) - D(a + b + c) + D(b + c) \geqq 0.$$

This completes the proof.

COROLLARY: *If* L *is a lattice, then* $D(a)$ *is a dimension function if and only if* $\Delta_D(a, ab, b) = 0$ *and* $\delta_D(a, b) > 0$ *for every* $a, b \in L$ *with* $a \neq b$.

PROOF: This follows immediately from Lemmas 1.2 (1) and 1.1 (6).

THEOREM 1.1: *Let* $D(a)$ *be a real-valued function defined on a lattice* L. *Then the statements,*

(1) $D(a)$ *is a norm for* L.
(2) $D(a)$ *is properly monotone increasing and satisfies*

$$D(a + b) + D(c) \leqq D(a) + D(b + c) \text{ for } c \leqq a, \tag{1.3}$$

(3) $D(a)$ *is an affine dimension function,*
are equivalent.

PROOF: We note that $D(a)$ is a norm for L if and only if $\delta_D(a, b) > 0$ for $a \neq b$ and $\Delta_D(a, b, c) \geqq 0$ for every $a, b, c \in L$. Then by Lemmas 1.1 (6) and 1.2 (5) the statements (1) and (2) are equivalent. To show the equivalence of (2) and (3) it suffices to show that the condition (1.2) is equivalent to (1.3) provided that $D(a)$ is monotone increasing. Suppose (1.2) holds and consider $a, b, c \in L$ with $c \leqq a$. By (1.2) we have

$$D(a) + D(b + c) \geqq D(a(b + c)) + D(a + b + c) \geqq D(c) + D(a + b).$$

Thus if $D(a)$ is monotone increasing, condition (1.2) implies (1.3). Now let (1.3) hold, and consider $a, b \in L$. Define $c \equiv ab$. Clearly $c \leqq a$, and (1.3) yields $D(a + b) + D(ab) \leqq D(a) + D(b + ab) = D(a) + D(b)$. Thus condition (1.3) implies (1.2). This completes the proof.

One of the striking differences between the theories of affine and projective geometry, when viewed from the lattice standpoint, is that in the former the modular law,

$$(a + b)c = a + bc \text{ for every } a \leqq c, \tag{1.4}$$

may fail for certain pairs b, c. The following theorem indicates the intimate relationship between modularity and the norm $D(a)$.

THEOREM 1.2: *If $D(a)$ is an affine dimension function defined on a lattice L then* (1.4) *holds for a given pair $b, c \in L$ whenever*

$$D(b + c) + D(bc) = D(b) + D(c).$$

PROOF: Let the pair $b, c \in L$ satisfy the condition of the theorem, and consider $a \leqq c$. Clearly $(a + b)c \geqq a + bc$. Hence by Lemma 1.1 (5) it suffices to show that $D(a + bc) \geqq D((a + b)c)$. Using Lemma 1.2 (4) (iii) we obtain

$$\begin{aligned} D((a + b)c) - D(a + bc) &\leqq D(a + b) + D(c) - D(a + b + c) - D(a + bc) \\ &= D(a + b) + D(c) - D(b + c) - D(a + bc) \\ &= D(a + b) + D(bc) - D(b) - D(a + bc) \\ &\leqq 0. \end{aligned}$$

This completes the proof.

Let us now consider a lattice L and a function $\delta(a, b)$ which is a metric for L. We investigate the possibility of realizing δ as the δ_D for a suitable function $D(a)$.

LEMMA 1.3: *If δ has the property $\Delta(a, b, c) = 0$ for $a \leqq b \leqq c$, then*

$$\text{l.u.b. } [\delta(c, b); c \leqq a] = \text{l.u.b. } [\delta(c, b); c \leqq b]$$

for every $a, b \in L$ with $a \leqq b$.

PROOF: Clearly it suffices to show that corresponding to an element $c \leqq b$ there is an element $c_1 \leqq a$ such that $\delta(c_1, b) \geqq \delta(c, b)$. Define $c_1 \equiv ca$. Hence $c_1 \leqq a$, and $c_1 \leqq c \leqq b$. Thus $\delta(c_1, b) = \delta(c_1, c) + \delta(c, b) \geqq \delta(c, b)$, and the lemma is proved.

LEMMA 1.4: *If φ_1 and φ_2 are real-valued monotone decreasing (increasing) functions defined on a sublattice L_0 of L, then*

$$\text{l.u.b.}[\varphi_1(a) + \varphi_2(a); a \in L_0] = \text{l.u.b. } [\varphi_1(a); a \in L_0] + \text{l.u.b. } [\varphi_2(a); a \in L_0].$$

PROOF: The argument used to prove the analogous fact about real functions of a real variable applies here.

THEOREM 1.3: *If δ has the property*

$$\text{l.u.b. } [\delta(a, b); b \leqq a] < + \infty \text{ for every } a \in L,$$

then there exists a norm $D(a)$ such that $\delta = \delta_D$ if and only if

$$(1)\ \Delta(a, b, c) = 0 \text{ for } a \leqq b \leqq c,$$

and

$$(2)\ \Delta(a, a + b, b) = 0 \text{ for every } a, b \in L.$$

The norm $D(a)$ is unique up to an additive constant.

PROOF: That (1) and (2) are necessary follows from parts (1) and (4) of Lemma 1.1. To prove the sufficiency define $D(a) \equiv$ l.u.b. $[\delta(c, a); c \leqq a]$. Using (1) and (2) one easily obtains for $c \leqq ab$

$$\delta(a, b) = 2\delta(c, a + b) - \delta(c, a) - \delta(c, b).$$

An application of Lemma 1.4 yields

$$\delta(a, b) + \text{l.u.b. } [\delta(c, a); c \leqq ab] + \text{l.u.b. } [\delta(c, b); c \leqq ab]$$
$$= 2 \text{ l.u.b. } [\delta(c, a + b); c \leqq ab].$$

The relation $\delta = \delta_D$ now follows from the definition of D and Lemma 1.3. To prove the uniqueness consider two norms D and E such that $\delta = \delta_D = \delta_E$. If $a \leqq b$, then $\delta(a, b) = D(b) - D(a) = E(b) - E(a)$. Thus $D(b) - E(b) = D(a) - E(a)$ if $a \leqq b$. Now let a, b be arbitrary. Clearly $ab \leqq a$ and $ab \leqq b$, whence $D(b) - E(b) = D(ab) - E(ab) = D(a) - E(a)$. This completes the proof.

COROLLARY: *If $0 \in L$ exists, then $D(a)$ may be chosen as $\delta(0, a)$, and then $a \leqq b$ if and only if $\Delta(0, a, b) = 0$.*

PROOF: When $0 \in L$ exists, we have by Lemma 1.3 l.u.b. $[\delta(c, a); c \leqq a] = \delta(0, a)$. If $a \leqq b$, then $\Delta(0, a, b) = 0$ by (1) since $0 \leqq a \leqq b$. From (1) and (2) we obtain

$$\begin{aligned} 2\delta(a + b, b) &= \delta(0, a + b) - \delta(0, b) + \delta(a + b, b) \\ &= \delta(0, a) + \delta(a, a + b) + \delta(a + b, b) - \delta(0, b) \\ &= \delta(0, a) + \delta(a, b) - \delta(0, b) \\ &= 2\Delta(0, a, b). \end{aligned}$$

Thus $\Delta(0, a, b) = 0$ yields $\delta(a + b, b) = 0$, and we have $a + b = b$, whence $a \leqq b$.

2. Continuity and Completeness of a Normed Lattice

In this section we shall be interested for the most part in the relation between the properties of continuity and completeness of a given normed additive system L. We shall show that completeness of L in the metric $\delta = \delta_D$ implies that L is continuous. A necessary and sufficient condition that L be complete, under the assumption that L is continuous, will be given at the close of the section. The usual concepts and notations of the theory of metric spaces, such as "fundamental sequence," "completeness," etc., will be assumed.[5]

[5] F. Hausdorff, op. cit., pp. 94, ff.

LEMMA 2.1: *Every increasing (decreasing) sequence of elements of L is either finite or denumerable.*

PROOF: Since $D(a)$ is properly monotone increasing, our lemma follows immediately from the fact that every monotone sequence of real numbers is either finite or denumerable.

LEMMA 2.2: *Let P be a partially ordered set for which every increasing (decreasing) sequence is finite or denumerable. Then every increasing (decreasing) sequence of elements of P has a sum (product) if and only if every non-empty subset of P has a sum (product).*

PROOF: The converse implication is trivial. Suppose now that every increasing sequence of elements of P has a sum. If it is false that every non-empty subset of P has a sum, then there exists a non-empty subset S of P with the following properties:

(1) S has no sum.

(2) The power $\aleph$ of S is a minimum for subsets of P satisfying (1).

Let Ω be the least ordinal of power $\aleph$. It follows that Ω is infinite and is a limit ordinal and that there exists a system $(a_\alpha\,;\ \alpha < \Omega)$, $a_\alpha \in P$, such that $S = [a_\alpha\,;\ \alpha < \Omega]$. Clearly the sets $[a_\beta\,;\ \beta < \alpha < \Omega]$ have sums. Define $b_\alpha \equiv \sum [a_\beta\,;\ \beta < \alpha]$. The sequence $(b_\alpha\,;\ \alpha < \Omega)$ is increasing and hence may be written $(b_i\,;\ i = 1, 2, \cdots)$. By hypothesis this sequence has a sum b. It follows that $b \geqq b_\alpha$ for every $\alpha < \Omega$. Since Ω is a limit ordinal, we have $b \geqq a_\alpha$ for every $\alpha < \Omega$. Let $c \geqq a$ for every $a \in S$. Then $c \geqq a_\alpha$ for every $\alpha < \Omega$. In particular, $c \geqq a_\beta$ for every $\beta < \alpha < \Omega$. Thus $c \geqq \sum [a_\beta\,;\ \beta < \alpha] = b_\alpha$ for every $\alpha < \Omega$. Hence $c \geqq b$, and b is effective as a sum for S. This contradicts (1) and the lemma is proved for this case. The alternate reading is dual.

LEMMA 2.3: *If $(a_i\,;\ i = 1, 2, \cdots)$ is a convergent increasing (decreasing) sequence of elements of L, then* $\lim a_i = \sum [a_i\,;\ i = 1, 2, \cdots]$ $(\lim a_i = \prod [a_i\,;\ i = 1, 2, \cdots])$.

PROOF: Let $(a_i\,;\ i = 1, 2, \cdots)$ be an increasing sequence of elements of L, and let $\lim a_i = a$. If $i = 1, 2, \cdots$ is fixed and if $j \geqq i$, then

$$\delta(a + a_i, a) \leqq \delta(a + a_i, a_j) + \delta(a_j, a)$$

$$= \delta(a + a_i, a_j + a_i) + \delta(a_j, a) \leqq 2\delta(a_j, a)$$

by Lemma 1.2 (4). Then

$$0 \leqq \delta(a + a_i, a) \leqq \lim 2\delta(a_j, a) = 0;$$

therefore $\delta(a + a_i, a) = 0$, and thus $a + a_i = a$, whence $a \geqq a_i$. Suppose now that $c \geqq a_i$ for $i = 1, 2, \cdots$; then by Lemma 1.2 (4)

$$\delta(a + c, c) = \delta(a + c, a_i + c) \leqq \delta(a, a_i),$$

and

$$0 \leqq \delta(a + c, c) \leqq \lim \delta(a, a_i) = 0.$$

Consequently, $\delta(a + c, c) = 0$, and therefore $a + c = c$, whence $c \geqq a$. Thus $a = \sum [a_i\,;\ i = 1, 2, \cdots]$. Since multiplication is not assumed to have the

same properties as addition, a dual proof cannot be given for the alternate reading. Let $(a_i\,;\, i = 1, 2, \cdots)$ be a decreasing sequence, and suppose that $\lim a_i = a$. If $i = 1, 2, \cdots$ is fixed and if $j \geqq i$, then

$$\delta(a + a_i\,,\, a_i) = \delta(a + a_i\,,\, a_j + a_j) \leqq \delta(a,\, a_j)$$

by Lemma 1.2 (4). Hence

$$0 \leqq \delta(a + a_i\,,\, a_i) \leqq \lim \delta(a,\, a_j) = 0;$$

therefore $\delta(a + a_i\,,\, a_i) = 0$, whence $a \leqq a_i$. Suppose now that $c \leqq a_i$ for $i = 1, 2, \cdots$; then

$$\delta(a + c,\, a) \leqq \delta(a + c,\, a_i) + \delta(a_i\,,\, a) = \delta(a + c,\, a_i + c) + \delta(a_i\,,\, a) \leqq 2\delta(a_i\,,\, a),$$

whence $\delta(a + c,\, a) = 0$, and $c \leqq a$. Thus $a = \prod [a_i\,;\, i = 1, 2, \cdots]$.

THEOREM 2.1: *L is complete if and only if all increasing and decreasing fundamental sequences are convergent.*

PROOF: The forward implication is trivial. To prove the converse implication it suffices to show that every fundamental sequence has a convergent subsequence. Consider a fundamental sequence. We may choose a subsequence, which we write $(a_i\,;\, i = 1, 2, \cdots)$ such that $\delta(a_i\,,\, a_j) \leqq (\frac{1}{2})^i$ for $j \geqq i$. By Lemma 1.2 (4) and induction we obtain

$$(2.1)\quad \delta(a_i + \cdots + a_j,\, a_i + \cdots + a_k) \leqq \frac{2}{2^m} \quad (j \geqq i,\, k \geqq i,\, m \equiv \min(k, j)).$$

Thus the sequences $(a_i + \cdots + a_{i+k}\,;\, k = 1, 2, \cdots)$ are increasing fundamental sequences for each $i = 1, 2, \cdots$. By hypothesis there is a sequence $(b_i\,;\, i = 1, 2, \cdots)$ for which $b_i = \lim_k (a_i + \cdots + a_{i+k})$. Thus $b_i = \sum [a_j\,;\, j = i, i + 1, \cdots]$ by Lemma 2.3. Using (2.1) we obtain for $j \geqq i$

$$\delta(a_i + \cdots + a_j,\, a_i) = D(a_i + \cdots + a_j) - D(a_i) \leqq \frac{2}{2^i}.$$

By Lemma 1.1 (1), $\delta(b_i\,,\, a_i) = D(b_i) - D(a_i) \leqq 2/2^i$. Hence the sequence (b_i) is equivalent to the sequence (a_i). Therefore the decreasing sequence (b_i) is fundamental and consequently has a limit a. Clearly $a = \lim a_i$, and the proof is complete.

THEOREM 2.2: *If L is complete and if $D(a)$ is bounded both above and below, then L is a continuous lattice.*

PROOF: By Lemmas 2.2 and 2.3 it suffices to show that all increasing and decreasing sequences are convergent. Consider an increasing or decreasing sequence (a_i). The sequence of numbers $D(a_i)$, $i = 1, 2, \cdots$ is bounded and monotone, and hence fundamental. By Lemma 1.1 (1), $|\, D(a_i) - D(a_j)\,| = \delta(a_i\,,\, a_j)$. Thus the sequence (a_i) is fundamental and hence is convergent since L is complete. This completes the proof.

We shall turn from the principal object of this section to show how L may be isomorphically embedded in a continuous lattice. This has been done recently

by H. M. MacNeille.[6] He employed an extension of the notion of "Dedekind cuts" in the rational numbers. We find it more convenient to use only one of the two sets which constitute a "cut." We shall need the following definitions.

DEFINITION 2.1: For every subset S or L we define

$$S^+ \equiv [x; x \geqq y \text{ for every } y \in S],$$

$$S^- \equiv [x; x \leqq y \text{ for every } y \in S].$$

COROLLARY: (1) *If* $S, T \subset L$ *and* $S \subset T$, *then* $S^+ \supset T^+$, $S^- \supset T^-$

(2) *If* $S \subset L$, *then* $S^{+-+} = S^+$, $S^{-+-} = S^-$, $S \subset S^{+-}$, $S \subset S^{-+}$.

PROOF: Obvious.

DEFINITION 2.2: If $S \subset L$ then we say that S is *lower* (*upper*) in case $S^{+-} = S$ ($S^{-+} = S$) or equivalently, S is of the form T^- (T^+) with $T \subset L$. The set of all lower (upper) sets is denoted by $\mathfrak{L}$ ($\mathfrak{L}'$) (Elements of $\mathfrak{L}$ ($\mathfrak{L}'$) are denoted by $\mathfrak{a}, \mathfrak{b}, \mathfrak{c}$, etc. ($\mathfrak{a}', \mathfrak{b}', \mathfrak{c}'$, etc.)). Define $\mathfrak{L}_0$ as the class of all sets $[x; x \leqq a]$ for $a \in L$. Obviously $\mathfrak{L}_0 \subset \mathfrak{L}$.

COROLLARY 1: *If* $\mathfrak{L}_1 \subset \mathfrak{L}$ ($\mathfrak{L}_1' \subset \mathfrak{L}'$) *then* $\mathfrak{P}\mathfrak{L}_1 \in \mathfrak{L}$ ($\mathfrak{P}\mathfrak{L}_1' \in \mathfrak{L}'$).

PROOF: For each $\mathfrak{a} \in \mathfrak{L}_1$, $\mathfrak{a} \supset \mathfrak{P}\mathfrak{L}_1$, whence $\mathfrak{a} = \mathfrak{a}^+ \supset (\mathfrak{P}\mathfrak{L}_1)^{+-}$. Thus $\mathfrak{P}\mathfrak{L}_1 \supset (\mathfrak{P}\mathfrak{L}_1)^{+-}$, and it follows that $\mathfrak{P}\mathfrak{L}_1 = (\mathfrak{P}\mathfrak{L}_1)^{+-}$, i.e., $\mathfrak{P}\mathfrak{L}_1 \in \mathfrak{L}$. Similarly $\mathfrak{P}\mathfrak{L}_1' \in \mathfrak{L}'$.

COROLLARY 2: *If* $\mathfrak{a} \in \mathfrak{L}$, *then* $\mathfrak{a}$ *is an ideal, i.e.,* (a) $a, b \in \mathfrak{a}$ *implies* $a + b \in \mathfrak{a}$, *and* (b) $a \in \mathfrak{a}$, $b \leqq a$ *implies* $b \in \mathfrak{a}$. *If* $\mathfrak{a}' \in \mathfrak{L}'$ *then* $\mathfrak{a}'$ *is a dual ideal, i.e.,* ($\mathfrak{a}'$) $a, b \in \mathfrak{a}'$ *implies* $ab \in \mathfrak{a}'$, *and* (b') $a \in \mathfrak{a}'$, $b \geqq a$ *implies* $b \in \mathfrak{a}'$.

PROOF: Obvious.

If for $\mathfrak{a}, \mathfrak{b} \in \mathfrak{L}$ we define $\mathfrak{a} < \mathfrak{b}$ as $\mathfrak{a} \subset \mathfrak{b}$, $\mathfrak{a} \neq \mathfrak{b}$, then $\mathfrak{L}$ is a continuous lattice in which $\prod S = \mathfrak{P}S$, $\sum S = (\mathfrak{S}S)^{+-}$ for $S \subset \mathfrak{L}$. The subset $\mathfrak{L}_0$ of $\mathfrak{L}$ is a sublattice of $\mathfrak{L}$ which is lattice-isomorphic to L. We shall refer to $\mathfrak{L}$ as the *lattice-completion* of L. In similar fashion $\mathfrak{L}'$ is a continuous lattice; it is anti-isomorphic to $\mathfrak{L}$ under the correspondence $\mathfrak{a} \to \mathfrak{a}^+$.

The notation $\delta(S, T)$, where $S, T \subset L$ and $S, T \neq \Theta$, means[7] g.l.b. $[\delta(a, b); a \in S, b \in T]$.

DEFINITION 2.3: We say that L satisfies B_1 in case $\delta(a, \mathfrak{a}^+) = 0$ for every $\mathfrak{a} \in \mathfrak{L}$ for which $\mathfrak{a}, \mathfrak{a}^+ \neq \Theta$.

DEFINITION 2.4: We say that L satisfies B_2 in case the following two statements hold:

(a) Let $(a_i, i = 1, 2, \cdots)$ be an increasing sequence. If $[a_i; i = 1, 2, \cdots]^+ = \Theta$, then $\lim D(a_i) =$ l.u.b. $[D(a); a \in L]$; if $[a_i; i = 1, 2, \cdots]^+ \neq \Theta$, then $\delta([a_i; i = 1, 2, \cdots], [a_i; i = 1, 2, \cdots]^+) = 0$.

(b) Let $(a_i; i = 1, 2, \cdots)$ be a decreasing sequence. If $[a_i; i = 1, 2, \cdots]^- = \Theta$, then $\lim D(a_i) =$ g.l.b. $[D(a); a \in L]$; if $[a_i; i = 1, 2, \cdots] \neq \Theta$, then $\delta([a_i; i = 1, 2, \cdots], [a_i; i = 1, 2, \cdots]^-) = 0$.

---

[6] *Partially ordered sets*, Trans. Amer. Math. Soc., 42 (1937), pp. 416–460.

[7] F. Hausdorff, op. cit., p. 145.

We shall see in Section 3, where these ideas will be extensively used, that if L satisfies B_2, then L satisfies B_1. It is easy to show by examples that the converse is not true.

THEOREM 2.3: *Let L be continuous, and suppose that $D(a)$ is bounded above and below. Then L is complete if and only if L satisfies B_2.*

PROOF: Since every increasing or decreasing sequence is fundamental when the function $D(a)$ is bounded, the necessity follows immediately from Lemma 2.3. To prove the sufficiency it suffices, by Theorem 2.1, to show that every increasing or decreasing fundamental sequence is convergent. Consider an increasing fundamental sequence $(a_i\,;\ i = 1, 2, \cdots)$. Since L is continuous $\sum [a_i\,;\ i = 1, 2, \cdots]$ exists, and $[a_i\,;\ i = 1, 2, \cdots]^+ \neq \Theta$. Clearly $a_i \leqq \sum [a_i] \leqq b$ for $i = 1, 2, \cdots$, and every $b \in [a_i]^+$. Thus by Theorem 1.3 $\delta(a_i, \sum [a_j]) \leqq \delta(a_i, b)$ for $i = 1, 2, \cdots$ and every $b \in [a_i]^+$. The fact that L satisfies B_2 then implies that $\sum [a_j]$ is effective as a limit of the sequence (a_i). Decreasing sequences may be treated in a similar way. This completes the proof.

3. Equivalence of the Extensions

The Cantor method of completing a metric space is well known.[8] A description of the lattice-completion of a given partially ordered set was given in Section 2. Thus our normed lattice L possesses two extensions, viz., the lattice-completion and the completion in the metric $\delta = \delta_D$; relations between these two extensions will be treated in this section. We shall show that the metric-completion $\mathfrak{M}$ of L is a continuous normed lattice. We shall also prove that when L satisfies B_1 the lattice-completion $\mathfrak{L}$ is a normed lattice, that $\mathfrak{L}$ is (metric) complete if and only if L satisfies B_2, and that the isomorphism of $\mathfrak{L}$ and $\mathfrak{M}$ is equivalent to the statement that L satisfies B_2.

We shall assume throughout this section that the norm $D(a)$ of our normed lattice L is bounded above and below. Roman capitals, A, B, C, etc., will denote elements of $\mathfrak{M}$, i.e., maximal classes of equivalent fundamental sequences. We note that since $D(a)$ is bounded, every increasing or decreasing sequence is fundamental.

LEMMA 3.1: *If $A \in \mathfrak{M}$, $(a_i) \in A$, then* $\lim D(a_i)$ *exists and depends on A only, and not on the choice of the sequence $(a_i) \in A$.*

PROOF: This follows immediately from Lemma 1.1 (2).

DEFINITION 3.1: If $A \in \mathfrak{M}$, we define $D(A)$ as the unique common value of $\lim D(a_i)$ for all $(a_i) \in A$.

LEMMA 3.2: *If $A, B \in \mathfrak{M}$, $(a_i) \in A$, $(b_i) \in B$, then* $\lim \delta(a_i, b_i)$ *exists and depends on A, B only, and not on the choice of the sequences $(a_i) \in A$, $(b_i) \in B$.*

PROOF: This result, which follows from the uniform continuity of the metric δ, is contained in the general theory of metric spaces.[8]

DEFINITION 3.2: If $A, B \in \mathfrak{M}$, we define $\delta(A, B)$ as the unique common value of $\lim \delta(a_i, b_i)$ for all $(a_i) \in A$, $(b_i) \in B$.

[8] F. Hausdorff, op. cit., pp. 106–107.

COROLLARY: *$\delta(A, B)$ is a metric for $\mathfrak{M}$.*

PROOF: This follows immediately from Lemma 3.2 and the fact that δ is a metric for L.

LEMMA 3.3: *If $A, B \in \mathfrak{M}$, $(a_i) \in A$, $(b_i) \in B$, then the equivalence class determined by the sequence $(a_i + b_i)$ belongs to $\mathfrak{M}$ and depends on A, B only, and not on the choice of the sequences $(a_i) \in A$, $(b_i) \in B$.*

PROOF: Let $A, B \in \mathfrak{M}$, $(a_i) \in A$, $(b_i) \in B$. By Lemma 1.2 (4) (iv) we obtain

$$\delta(a_i + b_i, a_j + b_j) \leqq \delta(a_i, a_j) + \delta(b_i, b_j);$$

whence the equivalence class determined by the sequence $(a_i + b_i)$ belongs to $\mathfrak{M}$. If $(a_i') \in A$, $(b') \in B$, then Lemma 1.2 (4) (iv) yields

$$\delta(a_i' + b_i', a_i + b_i) \leqq \delta(a_i', a_i) + \delta(b_i', b_i).$$

Consequently, the sequences $(a_i + b_i)$ and $(a_i' + b_i')$ determine the same equivalence class.

DEFINITION 3.3: If $A, B \in \mathfrak{M}$, we define $A + B$ as the unique equivalence class determined by the sequences $(a_i + b_i)$, where $(a_i) \in A$, $(b_i) \in B$. If $A, B \in \mathfrak{M}$, then $A < B$ means $A + B = B$ with $A \neq B$.

LEMMA 3.4: *If $A, B \in \mathfrak{M}$, then $\delta(A, B) = 2D(A + B) - D(A) - D(B)$.*

PROOF: This follows immediately from Definition 3.2 and the fact that $D(a)$ is a norm for L.

LEMMA 3.5: *$\mathfrak{M}$ is a complete normed additive system with norm $D(A)$.*

PROOF: The operation $A + B$ is evidently associative, commutative, and idempotent. To show that $\mathfrak{M}$ is a partially ordered set, it suffices to prove that $A \leqq B$, $B \leqq C$ implies $A \leqq C$, and that $A \leqq B$, $B \leqq A$ implies $A = B$. If $A \leqq B$, $B \leqq C$, clearly

$$A + C = A + (B + C) = (A + B) + C = B + C = C,$$

whence $A \leqq C$. If $A \leqq B$, $B \leqq A$ we have $B = A + B = B + A = A$. Thus $\mathfrak{M}$ is a partially ordered set. To show that $A + B$ is effective as a least upper bound for A and B we have

$$A + (A + B) = (A + A) + B = A + B.$$

Hence $A + B \geqq A$, and by symmetry $A + B \geqq B$. If $C \geqq A, B$, then

$$(A + B) + C = A + (B + C) = A + C = C,$$

and it follows that $C \geqq A + B$. Thus $\mathfrak{M}$ is an additive system. We see from Lemma 3.4 and Corollary Definition 3.2 that $D(A)$ is a norm for $\mathfrak{M}$. By the general theory of metric spaces, $\mathfrak{M}$ is complete.[8] Hence $\mathfrak{M}$ is a complete normed additive system.

THEOREM 3.1: *$\mathfrak{M}$ is a continuous normed lattice.*

PROOF: This is immediate by Lemma 3.5 and Theorem 2.2.

Let us now consider the lattice-completion $\mathfrak{L}$ of L.

DEFINITION 3.4: For every non-empty $\mathfrak{a} \in \mathfrak{L}$ ($\mathfrak{a}' \in \mathfrak{L}'$) we define $D(\mathfrak{a})$ ($D'(\mathfrak{a}')$) as

$$\text{l.u.b. } [D(a);\ a \in \mathfrak{a}] \quad (\text{g.l.b. } [D(a);\ a \in \mathfrak{a}']).$$

We define $D(\Theta) \equiv D'(L)$ if $\Theta \in \mathfrak{L}$, and $D'(\Theta) \equiv D(L)$ if $\Theta \in \mathfrak{L}'$.

COROLLARY: *If L satisfies B_1, then $D(\mathfrak{a}) = D'(\mathfrak{a}^+)$ for every $\mathfrak{a} \in \mathfrak{L}$, and conversely.*

PROOF: Obvious.

REMARK: By Corollary Definition 3.4 we see that condition B_1 means that the functions D and D' (defined over $\mathfrak{L}$ and $\mathfrak{L}'$ respectively) correspond under the anti-isomorphism $\mathfrak{a} \to \mathfrak{a}^+$. When B_1 holds we might, without altering the results, have defined as lattice-completion of L the set of "cuts," i.e., pairs $(\mathfrak{a}, \mathfrak{a}^+)$ defining $D((\mathfrak{a}, \mathfrak{a}^+))$ as the common value of $D(\mathfrak{a})$, $D'(\mathfrak{a}^+)$.

DEFINITION 3.5: If $\mathfrak{a}, \mathfrak{b} \in \mathfrak{L}$ we define $\delta(\mathfrak{a}, \mathfrak{b})$ as $2D(\mathfrak{a} + \mathfrak{b}) - D(\mathfrak{a}) - D(\mathfrak{b})$.

DEFINITION 3.6: If $\mathfrak{a} \in \mathfrak{L}$ ($\mathfrak{a}' \in \mathfrak{L}'$) then a sequence $(a_i\ ;\ i = 1, 2, \cdots)$ is said to *generate* $\mathfrak{a}$ ($\mathfrak{a}'$) in case $(a_i\ ;\ i = 1, 2, \cdots)$ is fundamental, $a_i \in \mathfrak{a}$ ($a_i \in \mathfrak{a}'$) for $i = 1, 2, \cdots$, and

$$\lim D(a_i) = D(\mathfrak{a}) \qquad (\lim D(a_i) = D'(\mathfrak{a}')).$$

COROLLARY: *Let $\mathfrak{a} \in \mathfrak{L}$ ($\mathfrak{a}' \in \mathfrak{L}'$). If $\mathfrak{a} \neq \Theta$ ($\mathfrak{a}' \neq \Theta$), then there exists an increasing (decreasing) sequence $(a_i\ ;\ i = 1, 2, \cdots)$ generating $\mathfrak{a}$ ($\mathfrak{a}'$).*

PROOF: Suppose that $\mathfrak{a} \in \mathfrak{L}$ and that $\mathfrak{a} \neq \Theta$. Then $D(\mathfrak{a}) = \text{l.u.b. } [D(a);\ a \in \mathfrak{a}]$, and there exists $(b_i\ ;\ i = 1, 2, \cdots)$ with $b_i \in \mathfrak{a}$ such that $\lim D(b_i) = D(\mathfrak{a})$. Define $a_i \equiv \sum [b_j\ ;\ j = 1, \cdots, i]$. Clearly $a_i \in \mathfrak{a}$ for $i = 1, 2, \cdots$, and the sequence (a_i) is increasing. We have $D(b_i) \leqq D(a_i) \leqq D(\mathfrak{a})$, whence $\lim D(a_i) = D(\mathfrak{a})$. Thus the increasing sequence (a_i) generates $\mathfrak{a}$. The alternate reading is similarly proved.

LEMMA 3.6: *If $(a_i\ ;\ i = 1, 2, \cdots)$ and $(b_i\ ;\ i = 1, 2, \cdots)$ are equivalent increasing (decreasing) sequences, then $[a_i\ ;\ i = 1, 2, \cdots]^{+-} = [b_i\ ;\ i = 1, 2, \cdots]^{+-}$ ($[a_i\ ;\ i = 1, 2, \cdots]^{-+} = [b_i\ ;\ i = 1, 2, \cdots]^{-+}$).*

PROOF: Suppose that $(a_i\ ;\ i = 1, 2, \cdots)$, $(b_i\ ;\ i = 1, 2, \cdots)$ are equivalent increasing sequences. Define $S \equiv [a_i\ ;\ i = 1, 2, \cdots]$, $T \equiv [b_i\ ;\ i = 1, 2, \cdots]$. Let $x \in S^+$ i.e., $x \geqq a_i$ for $i = 1, 2, \cdots$. If $i = 1, 2, \cdots$ is fixed and $j \geqq i$, then

$$\begin{aligned}\delta(x + b_i, x) &\leqq \delta(x + b_i + a_j, x + b_i + b_j) + \delta(x + b_j, x + a_j)\\ &\leqq 2\delta(a_j, b_j)\end{aligned}$$

by Lemma 1.2. Thus

$$0 \leqq \delta(x + b_i, x) \leqq \lim 2\delta(a_j, b_j) = 0,$$

and it follows that $\delta(x + b_i, x) = 0$, whence $x \geqq b_i$, i.e., $x \in T^+$. We have proved that $S^+ \subset T^+$; by symmetry $T^+ \subset S^+$, whence $S^+ = T^+$, and $S^{+-} = T^{+-}$. Suppose now that $(a_i\ ;\ i = 1, 2, \cdots)$, $(b_i\ ;\ i = 1, 2, \cdots)$ are equivalent decreasing sequences. Let $x \in S^-$, i.e., $x \leqq a_i$ for $i = 1, 2, \cdots$. If $i = 1, 2, \cdots$

is fixed and $j \geqq i$, then

$$\delta(x + b_i, b_i) \leqq \delta(x + b_i + b_j, x + b_i + a_j) + \delta(b_i + a_j, b_i + b_j)$$
$$\leqq 2\delta(b_j, a_j)$$

by Lemma 1.2. Thus

$$0 \leqq \delta(x + b_i, b_i) \leqq \lim 2\delta(b_j, a_j) = 0,$$

and it follows that $x \leqq b_i$ i.e., $x \in T^-$. This and a symmetrical argument prove that $S^- = T^-$, whence $S^{-+} = T^{-+}$.

LEMMA 3.7: *Let* $\mathfrak{a} \in \mathfrak{L}$ $(\mathfrak{a}' \in \mathfrak{L}')$. *If* $(a_i\,;\, i = 1, 2, \cdots)$ $(b_i\,;\, i = 1, 2, \cdots)$ *generate* $\mathfrak{a}$ $(\mathfrak{a}')$, *then* $(a_i\,;\, i = 1, 2, \cdots)$ *and* $(b_i\,;\, i = 1, 2, \cdots)$ *are equivalent.*

PROOF: Let the sequences (a_i), (b_i) generate $\mathfrak{a}$. Then

$$D(a_i) \leqq D(a_i + b_i) \leqq D(\mathfrak{a}),$$

whence

$$D(\mathfrak{a}) = \lim D(a_i) \leqq \lim D(a_i + b_i) \leqq D(\mathfrak{a}),$$

and it follows from Lemma 1.1 (1) that $\lim \delta(a_i, a_i + b_i) = 0$. Consequently (a_i) is equivalent to the sequence $(a_i + b_i)$. By symmetry, the sequence (b_i) is equivalent to the sequence $(a_i + b_i)$, whence the sequences (a_i) and (b_i) are equivalent. The alternate reading is treated similarly.

LEMMA 3.8: *If an increasing (decreasing) sequence* $(a_i\,;\, i = 1, 2, \cdots)$ *generates* $\mathfrak{a}$ $(\mathfrak{a}')$ *then* $[a_i\,;\, i = 1, 2, \cdots]^{+-} = \mathfrak{a}$ $([a_i\,;\, i = 1, 2, \cdots]^{-+} = \mathfrak{a}')$.

PROOF: Consider first the case where $(a_i\,;\, i = 1, 2, \cdots)$ is an increasing sequence. Define $S \equiv [a_i\,;\, i = 1, 2, \cdots]$. We note that $S^{+-} \subset \mathfrak{a}$. Let Let $a \in \mathfrak{a}$, and define $b_i \equiv a_i + a$ for $i = 1, 2, \cdots$. Clearly the sequence (b_i) is increasing and $b_i \in \mathfrak{a}$ for $i = 1, 2, \cdots$. Now

$$D(\mathfrak{a}) \geqq \lim D(b_i) \geqq \lim D(a_i) = D(\mathfrak{a}).$$

Therefore the sequence (b_i) is equivalent to the sequence (a_i) and is hence fundamental; we have thus shown that the sequence (b_i) generates $\mathfrak{a}$. By Lemmas 3.7 and 3.6 we have $[b_i\,;\, i = 1, 2, \cdots]^{+-} = S^{+-}$. Clearly $a \in [b_i]^{+-}$; thus $\mathfrak{a} \subset S^{+-}$, whence $\mathfrak{a} = S^{+-}$, and the proof is complete in this case. The alternate reading is treated similarly.

REMARK: It is now easy to see from Corollary Definition 3.6 and Lemma 3.8 that if L satisfies B_2, then L satisfies B_1.

LEMMA 3.9: *Suppose that* L *satisfies* B_1. *If* $\mathfrak{a}$, $\mathfrak{b} \in \mathfrak{L}$, *and if the sequences* $(a_i\,;\, i = 1, 2, \cdots)$, $(b_i\,;\, i = 1, 2, \cdots)$ *are increasing and generate* $\mathfrak{a}$, $\mathfrak{b}$ *respectively, then the sequence* $(a_i + b_i\,;\, i = 1, 2, \cdots)$ *generates* $\mathfrak{a} + \mathfrak{b}$.

PROOF: Since the sequence $(a_i + b_i)$ is fundamental along with the sequences (a_i) and (b_i), it remains only to show that

$$\lim D(a_i + b_i) = D(\mathfrak{a} + \mathfrak{b}). \tag{3.1}$$

We prove first that $\mathfrak{b} \equiv [a_i + b_i ; i = 1, 2, \cdots]^{+-} = \mathfrak{a} + \mathfrak{b}$. Clearly

$$(\mathfrak{S}(\mathfrak{a}, \mathfrak{b}))^+ = \mathfrak{P}(\mathfrak{a}^+, \mathfrak{b}^+) \subset \mathfrak{b}^+.$$

By Lemma 3.8 $\mathfrak{a}^+ = [a_i ; i = 1, 2, \cdots]^+ \supset \mathfrak{b}^+$, and $\mathfrak{b}^+ = [b_i ; i = 1, 2, \cdots]^+ \supset \mathfrak{b}^+$. Hence $\mathfrak{P}(\mathfrak{a}^+, \mathfrak{b}^+) \supset \mathfrak{b}^+$. Thus $(\mathfrak{S}(\mathfrak{a}, \mathfrak{b}))^+ = \mathfrak{b}^+$, and therefore $\mathfrak{b} = \mathfrak{a} + \mathfrak{b}$. Now if either $\mathfrak{a}^+ = \Theta$ or $\mathfrak{b}^+ = \Theta$, then either $\mathfrak{a} = L$ or $\mathfrak{b} = L$ and the lemma is trivial. In the contrary case, since L satisfies B_1, there are sequences (c_j), (d_j) such that $c_j \in \mathfrak{a}^+$, $d_j \in \mathfrak{b}^+$ for $j = 1, 2, \cdots$, and $\lim \delta(c_i, a_i) = 0$, $\lim \delta(d_i, b_i) = 0$. It is clear that $c_j \geqq a_i$, $d_j \geqq b_i$ for $i, j = 1, 2, \cdots$. Let $x \in \mathfrak{a} + \mathfrak{b}$. Clearly $c_j + d_j \geqq x$, and

$$\delta(x + a_i + b_i, a_i + b_i) \leqq \delta(x + c_i + d_i + a_i + b_i, a_i + b_i)$$
$$\leqq \delta(c_i + d_i, a_i + b_i) \leqq \delta(a_i, c_i) + \delta(b_i, d_i).$$

Hence by Lemma 1.1 (1) we obtain

$$\lim D(x + a_i + b_i) = \lim D(a_i + b_i).$$

Therefore

$$D(\mathfrak{a} + \mathfrak{b}) \geqq \lim D(a_i + b_i) = \lim D(x + a_i + b_i) \geqq D(x)$$

for every $x \in \mathfrak{a} + \mathfrak{b}$, whence

$$D(\mathfrak{a} + \mathfrak{b}) \geqq \lim D(a_i + b_i) \geqq D(\mathfrak{a} + \mathfrak{b}).$$

Thus (3.1) holds, and the proof is complete.

THEOREM 3.2: *If L satisfies B_1, then $\mathfrak{L}$ is a normed lattice with norm $D(\mathfrak{a})$.*

PROOF: By Theorem 1.1 (2) it suffices to prove that $D(\mathfrak{a})$ is properly monotone increasing and that

$$D(\mathfrak{a} + \mathfrak{b}) + D(\mathfrak{c}) \leqq D(\mathfrak{a}) + D(\mathfrak{b} + \mathfrak{c}) \text{ for } \mathfrak{c} \leqq \mathfrak{a}. \tag{3.2}$$

Suppose that $\mathfrak{a} < \mathfrak{b}$. If $\mathfrak{a} \neq \Theta$, then $\mathfrak{b} \neq \Theta$, and there exist increasing sequences (a_i), (b_i) generating $\mathfrak{a}$ and $\mathfrak{b}$ respectively by Corollary Definition 3.4. Since $a_i \in \mathfrak{b}$ for $i = 1, 2, \cdots$, it follows that

$$D(a_i) \leqq \text{l.u.b. } [D(x); x \in \mathfrak{b}] = \lim D(b_i),$$

whence $\lim D(a_i) \leqq \lim D(b_i)$. If the equality sign holds, then the sequence (a_i) generates $\mathfrak{b}$, whence by Lemma 3.8 $\mathfrak{a} = [a_i, i = 1, 2, \cdots]^{+-} = \mathfrak{b}$, contrary to $\mathfrak{a} < \mathfrak{b}$. Let $\mathfrak{a} = \Theta$. Clearly $\mathfrak{b} \neq \Theta$, and $D(\mathfrak{b}) \geqq D(\mathfrak{a}) = \text{g.l.b. } [D(x); x \in L]$. Suppose that $D(\mathfrak{a}) = D(\mathfrak{b})$. If $b \in \mathfrak{b}$, then

$$D(\mathfrak{b}) \geqq D(b) \geqq \text{g.l.b. } [D(x); x \in L] = D(\mathfrak{b}),$$

and $D(b) = \text{g.l.b. } [D(x); x \in L]$. Now if $c \in L$ then $bc \in \mathfrak{b}$ (since $\mathfrak{b}$ is an ideal), whence also

$$D(bc) = \text{g.l.b. } [D(x); x \in L] = D(b).$$

But $bc \leqq b$, whence $bc = b$ and $b \leqq c$. Since $b \leqq c$ for every $c \in L$, it follows that $0 \in L$ exists and that $b = 0$. Consequently

$$b \in L^- = \Theta^{+-} = \mathfrak{a}^{+-} = \mathfrak{a}$$

contrary to $\mathfrak{a} = \Theta$. This proves that $D(\mathfrak{a})$ is properly monotone increasing. It remains only to establish (3.2). Clearly (3.2) holds if either $\mathfrak{a} = \Theta$ or $\mathfrak{b} = \Theta$. Hence let $\mathfrak{a}, \mathfrak{b} \neq \Theta$. Suppose first that $\mathfrak{c} \neq \Theta$. Then there exist increasing sequences (a_i'), (b_i) (c_i) generating $\mathfrak{a}$, $\mathfrak{b}$, $\mathfrak{c}$, respectively. Now define $a_i \equiv a_i' + c_i$. Then (a_i) clearly generates $\mathfrak{a}$, and $a_i \geqq c_i$ for $i = 1, 2, \cdots$. By Lemma 3.9 $(a_i + b_i)$ and $(b_i + c_i)$ generate $\mathfrak{a} + \mathfrak{b}$ and $\mathfrak{b} + \mathfrak{c}$ respectively. Since

$$D(a_i + b_i) + D(c_i) \leqq D(a_i) + D(b_i + c_i),$$

we have

$$\begin{aligned} D(\mathfrak{a} + \mathfrak{b}) + D(\mathfrak{c}) &= \lim D(a_i + b_i) + \lim D(c_i) \\ &\leqq \lim D(a_i) + \lim D(b_i + c_i) \\ &= D(\mathfrak{a}) + D(\mathfrak{b} + \mathfrak{c}). \end{aligned}$$

Finally let $\mathfrak{c} = \Theta$. Again, there exist increasing sequences (a_i), (b_i) generating $\mathfrak{a}$ and $\mathfrak{b}$ respectively, whence $(a_i + b_i)$ generates $\mathfrak{a} + \mathfrak{b}$. By Theorem 1.1 (3) and Definition 3.4 we obtain

$$\begin{aligned} D(\mathfrak{a} + \mathfrak{b}) + D(\Theta) &= \lim D(a_i + b_i) + \text{g.l.b. } [D(x); x \in L] \\ &\leqq \lim (D(a_i + b_i) + D(a_i b_i)) \\ &\leqq \lim D(a_i) + \lim D(b_i) = D(\mathfrak{a}) + D(\mathfrak{b} + \Theta). \end{aligned}$$

This completes the proof.

COROLLARY: *If L satisfies B_1 then $\mathfrak{L}_0$ (cf. Definition 2.2) and L are isometric.*

PROOF: The isomorphism $a \to [x; x \leqq a]$ between L and $\mathfrak{L}_0$ clearly preserves the norm D and hence preserves the metric δ.

THEOREM 3.3: *If L satisfies B_1 and if $\mathfrak{L}$ is (metric) complete, then L satisfies B_2.*

PROOF: This follows immediately from Theorem 3.2 and Lemma 2.3.

LEMMA 3.10: *Let L satisfy B_2. If $\mathfrak{a} \in \mathfrak{L}$ $(\mathfrak{a}' \in \mathfrak{L}')$, if $(a_i ; i = 1, 2, \cdots)$ is an increasing (decreasing) sequence, and if $[a_i ; i = 1, 2, \cdots]^{+-} = \mathfrak{a}$ $([a_i ; i = 1, 2, \cdots]^{-+} = \mathfrak{a}')$, then*

$$\lim D(a_i) = D(\mathfrak{a}) \ (\lim D(a_i) = D'(\mathfrak{a}')).$$

PROOF: Consider an increasing sequence (a_i), and define $S \equiv [a_i, i = 1, 2, \cdots]$. Then $S^{+-} = \mathfrak{a}$. If $S^+ = \Theta$, then $\mathfrak{a} = S^{+-} = L$ and $\lim D(a_i) = D(L) = D(\mathfrak{a})$. Hence let $S^+ \neq \Theta$. Since L satisfies B_2 we have $\delta(S, S^+) = \delta(S, \mathfrak{a}^+) = 0$. It follows that there exists a sequence (b_i) with $b_i \in \mathfrak{a}^+$ such that $\lim \delta(a_i, b_i) = 0$. Thus

$$0 = \lim (D(b_i) - D(a_i)) \geqq D(\mathfrak{a}) - \lim D(a_i) \geqq 0,$$

whence $\lim D(a_i) = D(\mathfrak{a})$. The alternate reading may be similarly proved.

THEOREM 3.4: *Let L satisfy B_2. Then $\mathfrak{L}$ is (metric) complete.*

PROOF: By Theorem 2.1 it suffices to show that every increasing or decreasing sequence in $\mathfrak{L}$ is convergent. Let $(\mathfrak{a}_i ; i = 1, 2, \cdots)$ be increasing, i.e., $\mathfrak{a}_i \leqq \mathfrak{a}_{i+1}$ for $i = 1, 2, \cdots$. For each $i = 1, 2, \cdots$ we may select an increasing sequence $(b'_{ik}; k = 1, 2, \cdots)$ which generates $\mathfrak{a}_i$. Then define $b_{ik} \equiv \sum [b'_{jk}; j = 1, \cdots, i]$; clearly $(b_{ik} ; k = 1, 2, \cdots)$ generates a_i, and $b_{ik} \leqq b_{i+1,k}$, $b_{ik} \leqq b_{i,k+1}$ for i, $k = 1, 2, \cdots$. Define $a_i \equiv b_{ii}$ and $S \equiv [a_i ; i = 1, 2, \cdots]$. It will be shown that $\mathfrak{a} \equiv \sum [\mathfrak{a}_i; i = 1, 2, \cdots] = S^{+-}$. Clearly $a_i \in \mathfrak{a}_i$, whence $S \subset \mathfrak{S} [\mathfrak{a}_i; i = 1, 2, \cdots] \subset \mathfrak{a}$. Hence $S^+ \supset \mathfrak{a}^+$. Now

$$\mathfrak{a}^+ = (\mathfrak{S}[\mathfrak{a}_i ; i = 1, 2, \cdots])^+ = \mathfrak{P}[\mathfrak{a}_i^+; i = 1, 2, \cdots].$$

If $x \in S^+$, i.e., $x \geqq a_i$ for $i = 1, 2, \cdots$, then $x \geqq b_{ik}$ for $i, k = 1, 2, \cdots$, whence $x \in [b_{ik} ; k = 1, 2, \cdots]^+ = \mathfrak{a}_i^+$ for $i = 1, 2, \cdots$, and $x \in \mathfrak{P} [\mathfrak{a}_i^+ ; i = 1, 2, \cdots]$. Therefore $S^+ \subset \mathfrak{a}^+$, and $S^+ = \mathfrak{a}^+$. Consequently $S^{+-} = \mathfrak{a}$, and by Lemma 3.10 $\lim D(a_i) = D(\mathfrak{a})$. Now $D(\mathfrak{a}_i) \geqq D(a_i)$, and $D(\mathfrak{a}) \geqq D(\mathfrak{a}_i)$ for $i = 1, 2, \cdots$, whence

$$\lim D(\mathfrak{a}_i) \geqq \lim D(a_i) = D(\mathfrak{a}) \geqq \lim D(\mathfrak{a}_i).$$

Thus $\lim D(\mathfrak{a}_i) = D(\mathfrak{a})$, i.e., $\lim \delta(\mathfrak{a}_i , \mathfrak{a}) = 0$, and the sequence $(\mathfrak{a}_i)$ is convergent to $\mathfrak{a}$. Let $(\mathfrak{a}_i ; i = 1, 2, \cdots)$ be decreasing, i.e., $\mathfrak{a}_i \geqq \mathfrak{a}_{i+1}$ for $i = 1, 2, \cdots$. For each $i = 1, 2, \cdots$ we may select a decreasing sequence $(b'_{ik}; k = 1, 2, \cdots)$ which generates $\mathfrak{a}_i^+$. Then define $b_{ik} \equiv \prod [b'_{jk} ; j = 1, \cdots, i]$; clearly $(b_{ik} ; k = 1, 2, \cdots)$ generates $\mathfrak{a}_i^+$, and $b_{ik} \geqq b_{i+1,k}$, $b_{ik} \geqq b_{i,k+1}$ for i, $k = 1, 2, \cdots$. Define $a_i \equiv b_{ii}$ and $T \equiv [a_i ; i = 1, 2, \cdots]$. It will be shown that $\mathfrak{b} \equiv \prod [\mathfrak{a}_i ; i = 1, 2, \cdots] = T^-$. Let $y \in T^-$, i.e., $y \leqq a_i$ for $i = 1, 2, \cdots$. Then $y \leqq b_{ik}$, i.e., $y \in [b_{ik} ; k = 1, 2, \cdots]^- = \mathfrak{a}_i$ for $i = 1, 2, \cdots$, whence $y \in \mathfrak{b}$. If $x \in \mathfrak{b}$, then $x \in [b_{ik} ; k = 1, 2, \cdots]^-$ for $i = 1, 2, \cdots$, whence $x \in T$. Thus $T^- = \mathfrak{b}$. By Lemma 3.10

$$\lim D(a_i) = D'(\mathfrak{b}^+) = D(\mathfrak{b}),$$

the last equality holding by Corollary Definition 3.4. Therefore

$$\lim D(a_i) \geqq \lim D(\mathfrak{a}_i) \geqq D(\mathfrak{b}) = \lim D(a_i),$$

and $\lim D(\mathfrak{a}_i) = D(\mathfrak{b})$, i.e., $\lim \delta(\mathfrak{a}_i , \mathfrak{b}) = 0$, and the sequence $(\mathfrak{a}_i)$ is convergent to $\mathfrak{b}$. This completes the proof.

REMARK: Since we have stated B_2 in terms of L, its meaning may need some clarification. It is easily shown in the light of Theorem 3.4 that B_2 is equivalent to the complete additivity and complete multiplicativity of D as defined on $\mathfrak{L}$, i.e., $\lim D(\mathfrak{a}_i) = D(\sum [\mathfrak{a}_i ; i = 1, 2, \cdots])$ if $(\mathfrak{a}_i ; i = 1, 2, \cdots)$ is increasing, and $\lim D(\mathfrak{a}_i) = D(\prod [\mathfrak{a}_i ; i = 1, 2, \cdots])$ if $(\mathfrak{a}_i ; i = 1, 2, \cdots)$ is decreasing.

LEMMA 3.11: *The lattice $\mathfrak{L}_0$ is everywhere dense in $\mathfrak{L}$.*

PROOF: Suppose first that $\mathfrak{a} = \Theta$, and let $(b_i ; i = 1, 2, \cdots)$ be a decreasing

sequence generating $\Theta^+ = L$. Then

$$\lim_i \delta([x; x \leqq b_i], \mathfrak{a}) = \lim_i (D([x; x \leqq b_i]) - D(\mathfrak{a}))$$
$$= \lim D(b_i) - D(\mathfrak{a})$$
$$= \lim D(b_i) - D'(L) = 0.$$

Suppose that $\mathfrak{a} \neq \Theta$, and let $(b_i\,;\ i = 1, 2, \cdots)$ be an increasing sequence generating $\mathfrak{a}$. Then

$$\lim \delta([x; x \leqq b_i], \mathfrak{a}) = \lim (D(\mathfrak{a}) - D([x; x \leqq b_i]))$$
$$= D(\mathfrak{a}) - \lim D(b_i) = 0.$$

Thus if $\mathfrak{a} \in \mathfrak{L}$ and $e > 0$, there is an element $\mathfrak{a}_0 \in \mathfrak{L}_0$ such that $\delta(\mathfrak{a}, \mathfrak{a}_0) \leqq e$, i.e., $\mathfrak{L}_0$ is everywhere dense in $\mathfrak{L}$.

THEOREM 3.5: *Let L satisfy B_2. Then there is a one-to-one correspondence between $\mathfrak{L}$ and $\mathfrak{M}$ preserving the lattice operations, the norm D and the metric δ.*

PROOF: Since $\mathfrak{L}$ is a metric-complete extension of L such that L (i.e., $\mathfrak{L}_0$) is everywhere dense in $\mathfrak{L}$, it follows from the general theory of metric spaces that there exists a one-to-one isometric correspondence $\mathfrak{a} \to A(\mathfrak{a})$ between $\mathfrak{L}$ and $\mathfrak{M}$, and that such a correspondence is defined in the following manner. If $\mathfrak{a} \in \mathfrak{L}$ select a sequence $(\mathfrak{a}_i) \in \mathfrak{L}_0$, i.e., a sequence $(a_i) \in L$ such that

$$\lim \delta(a_i, \mathfrak{a}) = \lim_i \delta([x; x \leqq a_i], \mathfrak{a}) = 0;$$

define $a_{ij} \equiv a_i$ for $i, j = 1, 2, \cdots$, and A_i as the equivalence class of $(a_{ij}\,;\ j = 1, 2, \cdots)$. Then the sequence (A_i) is fundamental in $\mathfrak{M}$ and therefore has a limit $A(\mathfrak{a}) \in \mathfrak{M}$; $A(\mathfrak{a})$ depends only on $\mathfrak{a}$, and the correspondence $\mathfrak{a} \to A(\mathfrak{a})$ has the desired properties. Indeed, for every sequence $(b_i) \in A(\mathfrak{a})$ we have

$$D(A(\mathfrak{a})) = \lim D(b_i) = D(\mathfrak{a}),$$

whence $\mathfrak{a} \to A(\mathfrak{a})$ preserves the norm D. Moreover, if $\mathfrak{a} \neq \Theta$ then $A(\mathfrak{a})$ is the unique equivalence class determined by sequences generating $\mathfrak{a}$. Finally, in order to prove that $\mathfrak{a} \to A(\mathfrak{a})$ is a lattice isomorphism, it suffices to show that $\mathfrak{a} \leqq \mathfrak{b}$ is equivalent to $A(\mathfrak{a}) \leqq A(\mathfrak{b})$. Suppose first that $\mathfrak{a} \neq \Theta$, and let $(a_i\,;\ i = 1, 2, \cdots)$, $(b_i\,;\ i = 1, 2, \cdots)$ be increasing sequences generating $\mathfrak{a}$, $\mathfrak{b}$ respectively. If $\mathfrak{a} \leqq \mathfrak{b}$, then $(a_i + b_i\,;\ i = 1, 2, \cdots)$ generates $\mathfrak{a} + \mathfrak{b} = \mathfrak{b}$ by Lemma 3.9, whence $A(\mathfrak{b}) = A(\mathfrak{a}) + A(\mathfrak{b})$, and $A(\mathfrak{a}) \leqq A(\mathfrak{b})$. Conversely, if $A(\mathfrak{a}) \leqq A(\mathfrak{b})$, i.e., if $A(\mathfrak{a}) + A(\mathfrak{b}) = A(\mathfrak{b})$, then by Lemma 3.8

$$\mathfrak{a} = [a_i\,;i = 1, 2, \cdots]^{+-} \subset [a_i + b_i\,;i = 1, 2, \cdots]^{+-} = [b_i\,;i = 1, 2, \cdots]^{+-} = \mathfrak{b}$$

whence $\mathfrak{a} \leqq \mathfrak{b}$. Suppose now that $\mathfrak{a} = \Theta$, and let $(a_i) \in A(\mathfrak{a})$, $(b_i) \in A(\mathfrak{b})$. Then

$$D(\mathfrak{a}) \leqq \lim D(a_i b_i) \leqq \lim D(a_i) = D(\mathfrak{a}),$$

and since

$$\delta(a_i + b_i, b_i) = D(a_i + b_i) - D(b_i) \leqq D(a_i) - D(a_i b_i)$$

by Theorem 1.1 (3), it follows that $\lim \delta(a_i + b_i, b_i) = 0$, whence $A(\mathfrak{a}) + A(\mathfrak{b}) = A(\mathfrak{b})$, i.e., $A(\mathfrak{a}) \leqq A(\mathfrak{b})$. This completes the proof.

The essential result of this section may be stated in the following

FUNDAMENTAL THEOREM: *Let L be a normed lattice with norm $D(a)$. If L satisfies B_1 and if $D(a)$ is bounded above and below, then the lattice-completion and the metric-completion of L are (lattice- and norm-) isomorphic if and only if L satisfies B_2.*

PROOF: This is immediate from Theorems 3.3 and 3.5.

4. An Example

In this section we shall discuss briefly a special normed lattice. The system obtained is an instance of systems investigated by one of us.[9]

Let L' be a lattice satisfying the following conditions:

(4.1) $0 \in L'$, $1 \in L'$ exist;

(4.2) there is defined on L' a dimension function $D(a)$;

(4.3) $D'(0) = 0$, $D'(1) = 1$;

(4.4) L' is complete in the metric $\delta \equiv \delta_{D'}$;

(4.5) L' has no points, i.e., $a \in L'$, $a > 0$ implies the existence of $c \in L'$ with $a > c > 0$.

By Theorems 1.2 and 2.2 L' is a modular continuous lattice. If $T \subset L'$ we write $\sum' T$ $(\prod' T)$ for the sum (product) of T. If $T = [a, b]$, we write $a \oplus b$ $(a \otimes b)$ for $\sum' T(\prod' T)$. We assume further that L' is complemented, i.e.,

(4.6) if $a \in L'$, there exists $x \in L'$ such that $a \oplus x = 1$, $a \otimes x = 0$. It may easily be shown that L' is a (not necessarily irreducible) continuous geometry.[10]

Let $q \in L'$ be any fixed element $\neq 0, 1$, and define $S \equiv [x; 0 < x \leqq q]$, $L \equiv L' - S$. Clearly L is a partially ordered set. Now it is easily seen that L is a continuous lattice in which for $T \subset L$, $\sum T = \sum' T$, $\prod T = \prod' T$ if $\prod' T \in L$, $\prod T = 0$ if $\prod' T \in S$. If we define $D(a) \equiv D'(a)$ for $a \in L$, then it follows immediately from Theorem 1.1 and (4.2) that $D(a)$ is a norm for L and that $\delta_D(a, b) = \delta_{D'}(a, b)$ for $a, b \in L$. Since L is a continuous normed lattice, it is evident that L satisfies B_1. To show that L does not satisfy B_2, select a solution $x \in L'$ of the equations $q \oplus x = 1$, $q \otimes x = 0$. Clearly $x \in L$, $x \neq 0$. Now define $x_1 \equiv x$. By (4.5) there exists $y \in L'$ such that $0 < y < x_1$. Moreover, (4.6) together with the modular law implies the existence of $z \in L'$ with $y \oplus z = x_1$, $y \otimes z = 0$. Thus $z \neq 0$, $D'(y) + D'(z) = D'(x_1)$, and hence either $2D'(y) \leqq D(x_1)$ or $2D'(z) \leqq D'(x_1)$. In the former case define $x_2 \equiv y$, and in the latter case, $x_2 \equiv z$. Proceeding thus inductively, we obtain a decreasing sequence $x_1 > x_2 > \cdots$ with $x_1 = x$, $x_i \neq 0$, $2D'(x_{i+1}) \leqq D'(x_i)$. Hence $2^{i-1}D'(x_i) \leqq D'(x_1)$, and therefore $\lim \delta(x_i, 0) = \lim D'(x_i) = 0$. If $x_i \in S$ for some $i = 1, 2, \cdots$, then $x_i = q \otimes x_i \leqq q \otimes x = 0$, whence $x_i = 0$, contrary to $x_i \neq 0$. Thus

[9] L. R. Wilcox, *Modularity in the theory of lattices*, to appear in Annals of Math.

[10] That is, L'satisties Axioms I–V in J. v. Neumann, *Continuous geometry*, Proc. Nat. Acad. Sci., 22 (1936) (pp. 94–96).

$x_i \in L$ for $i = 1, 2, \cdots$. Let y be any element of S. Clearly $y \oplus x_i \in L$ for $i = 1, 2, \cdots$, and the sequence $(y \oplus x_i)$ is decreasing. If $z \in L$, $z \leqq y \oplus x_i$ for $i = 1, 2, \cdots$, then by Lemma 2.3

$$z \leqq \prod' [y \oplus x_i ; i = 1, 2, \cdots] = \lim_i (y \oplus x_i) = y \leqq a,$$

and it follows that $z = 0$. Therefore $\prod [y \oplus x_i; i = 1, 2, \cdots] = 0$. However, since

$$\delta(0, y \oplus x_i) = D(y \oplus x_i) \geqq D'(y) > 0,$$

we see that L does not satisfy B_2. In fact, we have shown that L is everywhere dense in L', since a given element $y \in S$ is the limit of a sequence $(y \oplus x_i)$ of elements of L. Thus the metric-completion $\mathfrak{M}$ of L is isomorphic to L', while L is its own lattice-completion.

Discussion of further examples will be reserved for a subsequent publication by one of us.

THE INSTITUTE FOR ADVANCED STUDY

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

EVALUATIONS OVER RESIDUATED STRUCTURES

BY MORGAN WARD AND R. P. DILWORTH

(Received June 6, 1938)

I. INTRODUCTION

1. In previous papers,[1] we have developed a theory of structures over which auxiliary operations of multiplication and residuation may be defined with properties analogous to the like-named operations in polynomial ideal theory. We have applied our results to generalize the various decomposition theorems of ideal theory (van der Waerden [1]) to extensive classes of structures.[2]

We give here some results on the evaluations of such structures. By an evaluation we mean a homomorphism (Ore [1], Chapter II) between the structure and a set of real numbers ordered by the relation $\leqq$ under which multiplication in the structure corresponds to addition of the reals. If we are willing to assume that the structure homomorphism preserves residuation, we can obtain results of a simplicity and finality comparable with the classic evaluation theory for domains of integrity and fields. (van der Waerden [1], Albert [1]). But this assumption unduly restricts the kinds of structures which may be evaluated and complicates the arithmetical interpretation of the "discrete evaluations" (part III of paper) which are our main concern. We shall accordingly assume it only incidentally. (Part IV of paper.)

2. Our main result is an arithmetical characterization of all discrete evaluations of a residuated structure with the ascending chain condition in terms of certain chains of primary elements belonging to the structure. No appeal is necessary to the Dedekind modular axiom or to the special decomposition theorems which ensue on assuming that every irreducible is primary (Ward-Dilworth [1], [2]).

We also discuss briefly some interesting topological questions suggested by the evaluation.

---

[1] See the references Ward-Dilworth [1], Dilworth [1], Ward [1], Ward [2] at the close of the paper. The idea goes back to Dedekind (Dedekind [1], but our only immediate predecessor seems to have been W. Krull (Krull [1]). We have expanded and elaborated the results summarized in Ward-Dilworth [1] in a paper "Residuated Lattices" (Ward-Dilworth [2]) which has been submitted for publication elsewhere. We take this occasion to correct some errors in Ward-Dilworth [1] section 4, page 163. In condition D 1, the exponent r should be one. In the fourth theorem, the words "and sufficient" should be struck out. The fifth theorem should be struck out in toto. We may add that we have greatly extended the results of this section in Ward-Dilworth [2] and largely freed them of their dependence on the modular axiom.

[2] Distributive structures are studied in detail in the paper Ward [2] in this journal.

The following examples show that evaluations are of frequent occurrence. Here πx denotes the evaluation function, and $\mathfrak{S}$ the basic structure.

(i) $\mathfrak{S}$, rational integers with union and cross-cut G.C.D. and L.C.M. and multiplication ordinary multiplication. $\pi a = 0$, a odd; $\pi a = 1$, a even, defines an evaluation.

(ii) $\mathfrak{S}$, ideals of a principal ideal ring, multiplication ordinary multiplication. Evaluation is essentially the same as the ordinary evaluation.

(iii) $\mathfrak{S}$, Boolean algebra or distributive structure, multiplication identified with cross-cut (Ward [2]). $\mathfrak{p}$ any prime ideal of the structure (Stone [1], Birkhoff [1]). $\pi a = 1$ if a is in $\mathfrak{p}$, $\pi a = 0$ otherwise, defines an evaluation.

(iv) $\mathfrak{S}$ a chain structure of finite length n:

$$i = a_1 > a_2 > \cdots > a_n .$$

Let $n = n_1 + n_2 + \cdots + n_l$ be any fixed partition of n into positive summands. Effect a class separation of $\mathfrak{S}$, $\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}_1 + \cdots + \mathfrak{S}_l$ by the rule $a_i \in \mathfrak{S}_k$ if $n_1 + n_2 + \cdots + n_{k-1} < i \leqq n_1 + n_2 + \cdots + n_k$. Define a multiplication over $\mathfrak{S}$ as follows. Let $b_k = a_{n_1+n_2+\cdots+n_k}$, $(k = 1, 2, \cdots, l)$. Then if $a_u \in \mathfrak{S}_i$ and $a_v \in \mathfrak{S}_j$, $a_u a_v = b_{i+j}$ if $i + j \leqq l$, $a_u a_v = b_l = a_n$ if $i + j \geqq l$. Then $\pi a = k$ if $a \in \mathfrak{S}_k$ is an evaluation.

(v) $\mathfrak{S}$, finite arithmetical lattice. Since such a lattice is a direct cross-cut of chains of finite length, the procedure of (iv) allows us to construct evaluations at will. The case when each $n_i = 1$ in (iv) gives the ordinary evaluations. One of us plans to discuss the residuation of such a lattice elsewhere.

(vi) $\mathfrak{S}$, an arbitrary chain structure. $a_1 \supset a_2 \supset a_3 \supset \cdots$ any selection of elements of $\mathfrak{S}$ where each a_i properly divides a_{i+1}. Define $\mathfrak{S}_i$ as the set of all elements x of $\mathfrak{S}$ such that $a_{i-1} \supset x \supset a_i$, $x \neq a_i$. Define a multiplication by the rule if $x \in \mathfrak{S}_i$, $y \in \mathfrak{S}_j$ then $xy = a_{i+j}$. The evaluation is defined then by $\pi x = k$ if $x \in \mathfrak{S}_k$.

(vii) $\mathfrak{S}$, a residuated structure with ascending chain condition. p any prime of $\mathfrak{S}$. $\pi a = 1$ if $p \supset a$, $\pi a = 0$ otherwise defines an evaluation. (See part III of paper.) This example applies to the ideals of any commutative ring with chain condition, but no modular condition need be assumed.

3. We assumarize here the notations and definitions we shall employ. We denote our structures by German capitals $\mathfrak{S}$, $\mathfrak{S}_i$, $\mathfrak{S}'$, $\cdots$. The letters $\mathfrak{U}$, $\cdots$, $\mathfrak{Y}$ are reserved to denote sub-sets of elements of our basic structure $\mathfrak{S}$ which are not necessarily structures. We use small latin letters a, b, $\cdots$ for the elements of our structure, and write $x \in \mathfrak{X}$ $(x \in \mathfrak{S})$ for the set $\mathfrak{X}$ (the structure $\mathfrak{S}$) contains the element x. $x \supset y$ or $y \subset x$, $x \not\supset y$, $x = y$ denote as usual x divides y, x does not divide y, x equals y. We use (x, y) for union and $[x, y]$ for cross-cut, reversing Ore's usage. (Ore [1], Ward [1], [2].) We assume that $\mathfrak{S}$ has a unit element i dividing every other element. The null element z divisible by

every other element need not exist. Multiplication $x \cdot y$ or xy and residuation $x:y$ are one-valued operations on $\mathfrak{S}$ to $\mathfrak{S}$ defined by the following conditions:

R 1. $a, b \in \mathfrak{S}$ *implies* $a:b \in \mathfrak{S}$.
R 2. $a:b = i$ *if and only if* $a \supset b$.
R 3. $a \supset b$ *implies* $a:c \supset b:c$ *and* $c:b \supset c:a$.
R 4. $(a:b):c = (a:c):b$.
R 5. $[a, b]:c = [a:c, b:c]$.
R 6. $c:(a, b) = [c:a, c:b]$.

M 1. $a, b \in \mathfrak{S}$ *implies* $ab \in \mathfrak{S}$.
M 2. $a = b$ *implies* $ac = bc$.
M 3. $ab = ba$.
M 4. $(ab)c = a(bc)$.
M 5. $ai = a$.
M 6. $a(b, c) = (ab, ac)$.

In addition, the two operations are assumed to be interconnected by the formulas

(3.1) $\qquad a \supset (a:b)b$; if $a \supset xb$ then $a:b \supset x$.

(3.2) $\qquad ab:a \supset b$; if $y:a \supset b$ then $y \supset ab$.

A structure over which both a residuation and a multiplication may be defined satisfying R 1–M 6 and (3.1)–(3.2) will be said to be *residuated* (Ward-Dilworth [1]). We shall assume that the reader is familiar with the elementary properties of residuation and multiplication such as $a:b = a:(a, b) = [a, b]:b$; $a:bc = (a:b):c$; $a:(a:b) \supset (a, b)$ and so on. (Ward [1], [2], Dilworth [1], Ward-Dilworth [2].)

If $\mathfrak{X}$, $\mathfrak{Y}$ are any two subsets of $\mathfrak{S}$, we denote by $(\mathfrak{X}, \mathfrak{Y})$ $[\mathfrak{X}, \mathfrak{Y}]$ and $\mathfrak{X}\mathfrak{Y}$ the sets consisting respectively of all unions, cross-cuts or products of elements of $\mathfrak{X}$ with elements of $\mathfrak{Y}$ (Ward [2]). We write $\mathfrak{X} \supseteq \mathfrak{Y}$ if every element of $\mathfrak{Y}$ lies in $\mathfrak{X}$. For example, $\mathfrak{X} \supseteq \mathfrak{X}^2$ means $\mathfrak{X}$ is closed under multiplication. We use $\mathfrak{X} + \mathfrak{Y}$ for the set-theoretic sum of $\mathfrak{X}$ and $\mathfrak{Y}$.

4. An element p of $\mathfrak{S}$ is said to be a *prime* if $p \supset ab$ implies $p \supset a$ or $p \supset b$, and *primary* if $p \supset ab$, $p \not\supset a$ implies $p \supset b^t$ for some t. The following lemmas are true in any residuated structure in which the ascending chain condition holds. They are readily proved by transcribing their analogues for commutative ideal theory (von der Waerden [1] chapter 12) into the language of structure theory.

LEMMA 4.1. *If q is primary, there exists a prime p such that $p \supset q \supset p^t$.*

p will be said to *correspond* to q.

LEMMA 4.2. *Let q and p be elements of $\mathfrak{S}$ with the properties*

(α) $q \supset ab$ *and* $q \not\supset a$ *imply* $p \supset b$.
(β) $p \supset q$.
(γ) $p \supset b$ *implies* $q \supset b^t$ *for some* t.

Then q is primary, and p is the prime element corresponding to q.

The union of any set $\mathfrak{X}$ will on occasion be called the leader of $\mathfrak{X}$.

If $q = [a, b]$ implies $q = a$ or $q = b$, q will be said to be *irreducible*.

II. Evaluations

5. Any set of real numbers closed with respect to addition forms a residuated structure with respect to the division relation "less than or equal to." The union (α, β) and cross-cut $[\alpha, \beta]$ of two real numbers α and β are respectively their minimum and maximum, while their "product" $\alpha\beta$ and "residual" $\alpha{:}\beta$ are respectively $\alpha + \beta$ and $\alpha - (\alpha, \beta)$. (Dilworth [1]; Ward-Dilworth [1] section 2, first theorem; Ward [3].) If the set of real numbers is bounded above, they still form a residuated structure provided that we take the product $\alpha\beta$ equal to the least upper bound of the set whenever $\alpha + \beta$ is greater than it. We shall use the letter σ to denote the least upper bound of values. If no upper bound exists, we take $\sigma = +\infty$, the ideal null element of the structure of all the reals.

A function π on $\mathfrak{S}$ to such a set of reals is called an *evaluation* of $\mathfrak{S}$ if the following four conditions are satisfied:

E 1. *For every element a of $\mathfrak{S}$, πa is a uniquely determined real number.*
E 2. *$a = b$ implies $\pi a = \pi b$.*
E 3. (i) $\pi(a, b) = (\pi a, \pi b)$ *and* (ii) $\pi[a, b] = [\pi a, \pi b]$.
E 4. $\pi ab = (\pi a + \pi b, \sigma)$.

We shall call the real number πa the *value* of the structure element a.

If the values of πa are bounded above, we shall say that the evaluation is bounded. Since for any a, $a = ai$, $i \supset a$ and $a \supset b$ implies $a = (a, b)$, we have

$$\pi i = 0 \quad \cdot \quad \pi x \geqq 0,\ x \in \mathfrak{S}. \tag{5.1}$$

$$a \supset b \text{ implies } \pi a \leqq \pi b. \tag{5.2}$$

Since every evaluation is a homomorphism, we may define by means of the evaluation a congruence relation $x \equiv y \pmod{\pi}$ over $\mathfrak{S}$, elements being congruent if and only if they have the same values. This congruence relation has the usual properties; that is, it is an equivalence relation, and if $a \equiv b \pmod{\pi}$ then for any c,

$$(a, c) \equiv (b, c), \qquad [a, c] \equiv [b, c], \qquad ac \equiv bc \pmod{\pi}.$$

We call the elements congruent to i the units of $\mathfrak{S}$ modulo π. They form a dense residuated sub-structure of $\mathfrak{S}$. Moreover if u is any unit

$$au \equiv a \pmod{\pi}, \text{ every } a \text{ of } \mathfrak{S}.$$

6. Since we may think of an evaluation as a mapping of the structure onto the metric space of the real numbers, the question arises as to the connection of the evaluation with the topology of the structure. Structures have been topologized in several ways. For instance, Glivenko [1, 2] has shown the identity of normed structures with certain types of metric spaces. Stone [1] has shown that Boolean algebras are mathematically equivalent to locally

bicompact totally disconnected topological spaces, and H. Wallman [1] has similarly treated the topology of a distributive structure. In this connection one of us has found that much of Wallman's theory for a distributive structure holds in any structure over which a multiplication is defined.

If we attempt to introduce a metric into the structure by the use of the evaluation following the method of Glivenko, the difficulty arises that we may have $a \supset b$ properly with $\pi a = \pi b$. If we attempt to connect the evaluation with the topology of the structure by the method of Stone and Wallman, it is not clear what the map of a point (that is, a structure ideal) should be in terms of the evaluation. Our results in this direction are incomplete, and belong properly to the general question of the topology of residuated structures which one of us (R. P. Dilworth) will treat elsewhere.

III. Discrete Evaluations—Cleavages

7. If α is any positive value, then the elements of $\mathfrak{S}$ whose values are 0, α, 2α, 3α, $\cdots$ obviously form a multiplicatively closed sub-structure of $\mathfrak{S}$. A particularly interesting and important case occurs when this sub-structure coincides with $\mathfrak{S}$ itself. We shall call the evaluation then *discrete*. For a discrete evaluation, there is no loss in generality in taking the values to be $0, 1, 2, \cdots$, the set breaking off or not accordingly as the evaluation is bounded or unbounded. Let $\mathfrak{S}_k$ denote the set of elements of $\mathfrak{S}$ with values k. Then we have a set-theoretic separation of $\mathfrak{S}$

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}_1 + \mathfrak{S}_2 + \cdots + \mathfrak{S}_k + \mathfrak{S}_{k+1} + \cdots, \tag{7.1}$$

where each $\mathfrak{S}_i$ is a structure, and $\mathfrak{S}_0$ is multiplicatively closed. We denote the set $\mathfrak{S}_1 + \mathfrak{S}_2 + \cdots$ by $\mathfrak{S}'$, so that

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}'. \tag{7.2}$$

We shall now introduce the important notion of a *cleavage*[3] of a residuated structure. Let $\mathfrak{S}$ be a structure. From now on we assume explicitly

N 1. *$\mathfrak{S}$ is residuated.*

N 2. *The ascending chain condition holds in $\mathfrak{S}$.*

Definition of a Prime Cleavage. *A separation of $\mathfrak{S}$*

(7.3) $\mathfrak{S} = \mathfrak{U} + \mathfrak{V}$, $\mathfrak{U}$, $\mathfrak{V}$ *no elements in common, $\mathfrak{U}$ non-empty, is called a prime cleavage of $\mathfrak{S}$ provided that*

$$\mathfrak{U} \supset \mathfrak{U}^2, \qquad \mathfrak{U} \supset (\mathfrak{U}, \mathfrak{S}), \qquad \mathfrak{V} \supset (\mathfrak{V}, \mathfrak{V}). \tag{7.31}$$

Theorem 7.1. *Every cleavage of $\mathfrak{S}$ determines a prime, and with every prime is associated a cleavage.*

---

[3] The idea goes back to Krull [1]. For the special case when the multiplication of the structure is the cross-cut operation (but no chain condition is assumed, so that the cleavage does not necessarily define a structure element), the notion was applied to Boolean algebras by M. H. Stone [1], and extended by G. Birkhoff [2] to any distributive structure. The "primary cleavages" we introduce here seem to be new.

PROOF. By N 2 and (7.31)(iii), the union of $\mathfrak{V}$ exists, and lies in $\mathfrak{V}$ (Ore [1] §2). Denote it by p. *p cannot divide an element of* $\mathfrak{U}$. For then by (7.31) (ii) $p = (p, u) \in \mathfrak{U}$ contrary to (7.3). Assume that

$$p \supset ab, \qquad p \not\supset a, \qquad a, b \text{ in } \mathfrak{S}.$$

Then $ab \in \mathfrak{V}$, $a \in \mathfrak{U}$. If $b \in \mathfrak{U}$, $ab \in \mathfrak{U}$ by (7.31) (i), and $p \not\supset ab$. Hence $b \in \mathfrak{V}$ and $p \supset b$. Hence p is a prime by definition. The conceivable case $p = i$ is excluded because then since $i \supset x$, every x, $\mathfrak{V}$ would contain $\mathfrak{S}$ and $\mathfrak{U}$ be empty.

We observe that the separation (7.2) induced by the evaluation of $\mathfrak{S}$ is a cleavage. Since both $\mathfrak{U}$ and $\mathfrak{V}$ in (7.3) are easily shown to be structures, we shall use (7.2) henceforth to denote any prime cleavage. Conversely, given (7.2) with associated prime p, we may define a bounded evaluation over $\mathfrak{S}$ by $\pi x = 0$ if $p \not\supset x$; $\pi x = 1$ if $p \supset x$, $x \in \mathfrak{S}$. We have thus proved

THEOREM 7.2. *Every discrete evaluation of $\mathfrak{S}$ determines a prime, and every prime determines at least one evaluation.*

8. We shall next extend the notion of a cleavage so as to characterize the primary elements of $\mathfrak{S}$.

With the notation of the previous section, let

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}' \tag{7.2}$$

be a cleavage with associated prime p, so that

$$\mathfrak{S}_0 \supset \mathfrak{S}_0^2, \qquad \mathfrak{S}_0 \supset (\mathfrak{S}_0, \mathfrak{S}), \qquad \mathfrak{S}' \supset (\mathfrak{S}', \mathfrak{S}'), \tag{8.1}$$

while

$$p \in \mathfrak{S}'; x \in \mathfrak{S}' \text{ implies } p \supset x. \tag{8.11}$$

DEFINITION OF A PRIMARY CLEAVAGE. *A separation of* $\mathfrak{S}'$

(8.2) $\mathfrak{S}' = \mathfrak{U}^* + \mathfrak{V}$, $\mathfrak{U}^*$, $\mathfrak{V}$ *no elements in common, is called a primary cleavage of* $\mathfrak{S}$ *provided that*

(8.21) $\mathfrak{U}^* \supset \mathfrak{S}_0\mathfrak{U}^*$; $\mathfrak{V} \supset \mathfrak{S}'^k$ *for some positive integer* k; $\mathfrak{V} \supset (\mathfrak{V}, \mathfrak{V})$.

THEOREM 8.1. *The leader of $\mathfrak{V}$ is a primary element q of $\mathfrak{S}$ whose corresponding prime is p.*

PROOF. The result is trivial if $\mathfrak{U}^*$ is empty, as then $\mathfrak{V} = \mathfrak{S}'$. Assume henceforth that $\mathfrak{U}^*$ is non-empty. N 2 and (8.21) (iii) guarantee that the leader q exists and lies in $\mathfrak{V}$. Assume that $q \supset ab$, $q \not\supset a$; a, b in $\mathfrak{S}$. Then either $a \in \mathfrak{S}_0$ or $a \in \mathfrak{U}^*$. In either case $b \in \mathfrak{S}'$. For if $b \in \mathfrak{S}_0$, $ab \in \mathfrak{S}_0$ or $ab \in \mathfrak{U}^*$ by (8.1) (i) and (8.21) (i). But if $b \in \mathfrak{S}'$, then $p \supset b$ by (8.11). We have thus shown that

(α) $q \supset ab$ and $q \not\supset a$ imply $p \supset b$.

By (8.11) and (8.21) (ii), we have

(β) $p \supset b$. (γ) If $p \supset b$, then $q \supset b^k$.

Hence by lemma 4.2, q is primary, and p is the prime corresponding to q.

THEOREM 8.2. *Every primary element of $\mathfrak{S}$ determines a primary cleavage.*

PROOF. Let q be any primary, and let p be its corresponding prime. Then by lemma 4.1

$$p \supset q \supset p^k \tag{8.3}$$

where we may assume that $k > 1$. Let (7.2) as before be the cleavage associated with p. We now separate $\mathfrak{S}'$ into two disjoint classes $\mathfrak{U}^*$ and $\mathfrak{V}$ as in (8.2) by the rule $x \in \mathfrak{U}^*$ if $q \not\supset x$; $x \in \mathfrak{V}$ if $q \supset x$, x any element of $\mathfrak{S}'$. We shall show that (8.21) holds.

First if $b \in \mathfrak{S}'$, then $p \supset b$. Hence $p^k \supset b^k$, so that by (8.3), $q \supset b^k$ or $b^k \in \mathfrak{V}$. Hence (8.21) (ii) holds. Also, if $q \supset b$, $q \supset c$ then $q \supset (b, c)$ so that (8.21) (iii) holds. Finally, assume that $a \in \mathfrak{S}_0$ and $b \in \mathfrak{U}^*$. Then since $p \supset ab$, either $ab \in \mathfrak{U}^*$ or $ab \in \mathfrak{V}$. If $ab \in \mathfrak{V}$, then $q \supset ab$. Since $b \in \mathfrak{U}^*$, $q \not\supset b$. Hence since q is primary, $q \supset a^t$ for some t. Then by (8.3), $p \supset a^t$ or $p \supset a$, contradicting $a \in \mathfrak{S}_0$. Hence if $a \in \mathfrak{S}_0$ and $b \in \mathfrak{U}^*$, then $ab \in \mathfrak{U}^*$ giving (8.21) (i).

It is important to observe that the set $\mathfrak{U}^*$ is in general not a sub-structure of $\mathfrak{S}$.

9. We return now to the evaluation π and the associated separation of $\mathfrak{S}$ into residue classes:

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}_1 + \mathfrak{S}_2 + \cdots + \mathfrak{S}_k + \mathfrak{S}_{k+1} + \cdots \tag{7.1}$$

Let us define

$$\mathfrak{S}^* = \mathfrak{S}_1 + \mathfrak{S}_2 + \cdots + \mathfrak{S}_{k-1}, \qquad \mathfrak{S}'' = \mathfrak{S}_k + \mathfrak{S}_{k+1} + \cdots$$

so that

$$\mathfrak{S}' = \mathfrak{S}^* + \mathfrak{S}''.$$

Then it is easy to see that a primary cleavage is defined, for the conditions (8.2), (8.21) are all satisfied. Hence we may state

THEOREM 9.1. *Let π be a discrete evaluation of a residuated structure in which the ascending chain condition holds, and let* (7.1) *be the corresponding separation into residue classes modulo π. Then the leader of the substructure $\mathfrak{S}_k$ of all elements of $\mathfrak{S}$ with the value $k \geqq 1$ is a primary element $q^{(k)}$ of $\mathfrak{S}$ which divides all elements of $\mathfrak{S}$ with values $\geqq k$ and belongs to the prime element $p = q^{(1)}$ leading $\mathfrak{S}_1$.*

THEOREM 9.2. *The primaries $q^{(1)}, q^{(2)}, \cdots$ are all irreducible.*

PROOF. If q leads $\mathfrak{S}_k$ and $q = [a, b]$, then by E 3, either $\pi q = \pi a$ or $\pi q = \pi b$. Hence by Theorem 9.1, either $q \supset a$ or $q \supset b$. Hence $q = [a, b]$ implies $q = a$ or $q = b$ so that q is irreducible.

Although a primary cleavage can be associated with an arbitrary primary q, an evaluation cannot in general be determined having q as one of its leaders. For $\mathfrak{S}^*$ and $\mathfrak{S}''$ must both be structures and admit of special multiplication rules not required in our general definition.

To give a simple illustration, consider the non-modular structure $\mathfrak{L}$ pictured.

A residuation and multiplication may be defined over $\mathfrak{L}$ by the rule given in theorem 8.1 of Ward-Dilworth [2]; viz. $xy = y$ if $x = i$; $xy = x$ if $y = i$; $xy = z$,

otherwise. i is the unit and p is a prime while all other elements are primary. We may define three distinct evaluations over $\mathfrak{L}$ by the separations

(1) $\mathfrak{L}_0 = \{i\}; \mathfrak{L}_1 = \{p, a, b, c, d, e, z\}$.

(2) $\mathfrak{L}_0 = \{i\}; \mathfrak{L}_1 = \{p, a, d\}; \mathfrak{L}_2 = \{b, c, e, z\}$.

(3) $\mathfrak{L}_0 = \{i\}; \mathfrak{L}_1 = \{p, b, e\}; \mathfrak{L}_2 = \{a, c, d, z\}$.

Thus the primaries a and b both have evaluations associated with them. But no other primary determines an evaluation.

Consider c for example. Its primary cleavage is $\mathfrak{L}_0 = \{i\}$, $\mathfrak{L}^* = \{p, a, b, e, d\}$, $\mathfrak{L}'' = \{c, z\}$. But $\mathfrak{L}^*$ is not a structure, so that no evaluation is defined. Similar results hold for d, e and z.

Thus while any discrete evaluation of a structure satisfying N 1 and N 2 determines a chain of primaries all associated with the same prime, not every such chain determines an evaluation. On the other hand, every prime determines at least one evaluation; namely that determined by $\pi a = 1$ if $p \supset a$; $\pi a = 0$ if $p \not\supset a$.

If the evaluation π is bounded, so that $\mathfrak{S}$ separates into a finite number of residue classes modulo π,

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}_1 + \cdots + \mathfrak{S}_n,$$

then the evaluation may be defined in a manner strictly analogous to the evaluations of a finite principal ideal ring. Namely, let the leader of $\mathfrak{S}_n$ be q, and let p be its associated prime. For any other element a of $\mathfrak{S}$, there is then a least power of p such that $q \supset ap^s$. We then may define

$$\pi a = n - s \text{ if } q \supset ap^s, q \not\supset ap^{s-1}.$$

In particular, this definition applies to any discrete evaluation of a residuated lattice of finite order.

IV. Evaluations Preserving Residuation

10. We shall conclude by giving a few properties of bounded evaluations under which residuation is preserved. Consider an evaluation satisfying E 1, E 2 and

E 3 (i) $\pi(a, b) = (\pi a, \pi b)$.

E 5 $\pi a{:}b = \pi a{:}\pi b = \pi a - (\pi a, \pi b)$.

E 6 *The evaluation is bounded.*

Let σ as before denote the least upper bound of the values of π. If $\mathfrak{S}$ does not contain a null element z, we may adjoin z to $\mathfrak{S}$ without destroying the residuation by defining $z{:}z = i$, $zi = z$, $z{:}x = zx = z$, $x \neq i$, $x \in \mathfrak{S}$. (This fact is a special instance of theorem 8.2 of Ward-Dilworth [2].) Clearly $\pi z = \sigma$.

THEOREM 10.1. *If a, b are any two elements of $\mathfrak{S}$, then*

E 4. $\pi ab = (\pi a + \pi b, \sigma)$.

PROOF. Adjoin z to $\mathfrak{S}$. Then by E 5, $\pi z{:}\pi ab = \pi(z{:}ab) = \pi((z{:}a){:}b) = (\pi z{:}\pi a){:}\pi b$ or $\sigma - \pi ab = \sigma - \pi a - (\sigma - \pi a, \pi b)$. Hence since σ is finite, $\pi ab = \pi a + (\sigma - \pi a, \pi b)$, giving E 4.

Let λ denote the greatest lower bound of all positive values of π.

THEOREM 10.2. *If $\lambda = 0$, every real number in the interval $(0, \sigma)$ is a limit point of values. If $\lambda > 0$, then the evaluation is discrete.*

This theorem does not require the evaluation to be bounded. The first part of the theorem uses only E 4, and is true for any evaluation. But the second part of the theorem depends essentially on E 5, as may be shown by simple examples.

PROOF. If $\lambda = 0$, we may select a sequence of elements $a_1, a_2, \cdots, a_n, \cdots$ of $\mathfrak{S}$ such that $\alpha_n = \pi a_n > 0$, $\alpha_{n+1} \leqq \alpha_n$, $\lim \alpha_n = 0$. Since α_n is positive, for any positive β there exists an integer r_n such that $r_n\alpha_n < \beta \leqq (r_n + 1)\alpha_n$ for all sufficiently large n. Then if $b_n = a_n^{r_n}$, by E 4, $\pi b_n = r_n\alpha_n$ and $\lim \pi b_n = \beta$.

Suppose that $\lambda > 0$. *Then there exists an element l of $\mathfrak{S}$ such that $\pi l = \lambda$.* For otherwise, we may pick a sequence of elements x_i of $\mathfrak{S}$ such that $\pi x_n > \pi x_{n+1} > \lambda$; $\lim \pi x_n = \lambda$. Choose m so that $\pi x_m < 2\lambda$. Then by E 5, $\pi x_m{:}x_{m+1} = \pi x_m - \pi x_{m+1} < 2\lambda - \lambda < \lambda$. Hence $\pi x_m = \pi x_{m+1}$, giving a contradiction.

Let b be any other element of $\mathfrak{S}$ with a positive value πb. Then we can choose a positive integer r such that $r\lambda \leqq \pi b < (r + 1)\lambda$. Then by E 5 and E 4 $\pi b{:}l^r = \pi b - \pi(b, l^r) = \pi b - r\lambda < \lambda$. Hence $\pi b = r\lambda$, and the evaluation is discrete.

THEOREM 10.3. *If a discrete evaluation satisfies the conditions* E 1, E 2, E 3 (i) *and* E 5, *then it satisfies* E 3 (ii); *that is*

$$\pi[a, b] = [\pi a, \pi b] \qquad \text{for any elements } a, b \text{ of } \mathfrak{S}.$$

PROOF. We may assume that $\pi b \geqq \pi a \geqq 0$, so that we need only prove that $\pi[a, b] = \pi b$ if $\pi a \leqq \pi b$.

Since the evaluation is discrete, we may assume that $\pi a = r$, $\pi b = s$ where r and s are positive integers. Furthermore, there exists an element l of $\mathfrak{S}$ such that $\pi l = 1$.

LEMMA 1. $\pi[l^r, l^s] = \pi l^s$.

For since $r \leqq s$, $l^r \supset l^s$ so that $[l^r, l^s] = l^s$. The result now follows from E 2.

LEMMA 2. *If* $\pi a = \pi c$ *then* $\pi[a, b] = \pi[c, b]$.

For $\pi[a, b]:[c, b] = \pi[a:[c, b], b:[c, b]] = \pi a:[c, b]$ by E 2. Since $c \supset [c, b]$, $\pi c \leqq \pi[c, b]$. Hence $\pi a \leqq \pi[c, b]$, so that $\pi a:[c, b] = 0$ by E 5. Thus $\pi[a, b]:[c, b] = 0$. Hence by E 5, $\pi[a, b] \leqq \pi[c, l]$. Similarly, $\pi[c, b] \leqq \pi[a, b]$, giving the lemma.

The theorem now follows easily. For since $\pi a = \pi l^r$ and $\pi b = \pi l^s$ by E 4, the lemmas give

$$\pi[a, b] = \pi[l^r, b] = \pi[l^r, l^s] = \pi l^s = \pi b.$$

If the evaluation preserves residuation, the class separation

$$\mathfrak{S} = \mathfrak{S}_0 + \mathfrak{S}_1 + \mathfrak{S}_2 + \cdots \tag{7.1}$$

is subject to very stringent conditions; for if $a \in \mathfrak{S}_i$ and $b \in \mathfrak{S}_j$, then we must have $a:b \in \mathfrak{S}_0$ or $a:b \in \mathfrak{S}_{i-j}$ according as $i - j \leqq 0$.

The reader can easily show thereby that the structure discussed in section 9 admits of no such residuation. The interesting correspondence we have developed between primes and evaluations is thus destroyed.

CALIFORNIA INSTITUTE OF TECHNOLOGY,
PASADENA, CALIFORNIA

REFERENCES

A. A. ALBERT
1. *Modern Higher Algebra*, Chicago, 1937.

GARETT BIRKHOFF
1. *Rings of sets*. Duke Math. Jour., vol. 3 (1937), pp. 443–454.

R. DEDEKIND
1. *Gesammelte Werke*, vol. 2, Braunschweig (1931), paper XXI.

R. P. DILWORTH
1. *Abstract residuation over lattices*. Bulletin Am. Math. Soc., vol. 44 (1938) pp. 262–268.

V. GLIVENKO
1. *Géométrie des systèmes de choses normées*. Am. Jour. of Math., vol. 58 (1936), pp. 799–828.
2. *Contribution à l'étude des systèmes de choses normées*. Am. Jour. of Math., vol. 59 (1937), pp. 941–956.

W. KRULL
1. *Axiomatische Begründung der allgemein en Idealtheorie*. Erlanger Sitzungsberichte, vol. 36 (1924), pp. 47–63.

O. ORE
1. *On the foundations of abstract algebra*. These Annals, vol. 36 (1935), pp. 406–437.

M. H. STONE
1. *The theory of representations of Boolean algebras*. Trans. Am. Math. Soc., vol. 40 (1936), pp. 37–111.

B. L. VAN DER WAERDEN
1. *Moderne Algebra*, vol. 2. Berlin (1931).

H. WALLMAN
1. *Lattices and topological spaces*. These Annals, vol. 39, January 1938, pp. 112–126.

M. WARD AND R. P. DILWORTH

1. *Residuated lattices.* Proc. Nat. Acad. Sci., vol. 24 (1938), pp. 162–164.
2. *Residuated lattices.* To appear shortly in the Transactions of the Am. Math. Soc.

M. WARD

1. *Residuation in Structures over which a multiplication is defined.* Duke Math. Journal, vol. 3 (1937), pp. 627–636.
2. *Structure Residuation.* These Annals, vol. 39 (1938), pp. 558–568.
3. *Some arithmetical applications of residuation.* Am. Jour. of Math., vol. 59 (1937), pp. 921–926.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

ON CERTAIN POWER SERIES HAVING INFINITELY MANY ZERO COEFFICIENTS

BY M. S. ROBERTSON

(Received September 19, 1938)

1. Introduction

Let $f(z)$ be regular within the unit circle and have the power series

$$f(z) = z + \sum_{2}^{\infty} a_n z^n. \tag{1.1}$$

When $f(z)$ is univalent within the unit circle it is well known (1) that

$$a_n = O(n).$$

If the function is also odd, that is, if

$$f(z) + f(-z) \equiv 0$$

identically, or in other words, if

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod 2,$$

then in this case there exists (2) an absolute constant A greater than one for which

$$|a_n| < A \qquad \text{for all } n.$$

It would seem as though the vanishing of one half of the coefficients prevents the other half from being unbounded.

For the univalent functions with real coefficients the sharp result

$$|a_n| \leqq n \qquad \text{for all } n$$

is known (3). This fact was obtained by showing it to be also true for the larger class of analytic functions called by W. Rogosinski (4) typically-real functions, that is, those functions $f(z)$ which assume real values for, and only for, real values of z when z lies within the unit circle. However, for the class of *odd* typically-real functions the coefficients are not bounded, though they are bounded for odd univalent functions, a sub-class when the coefficients are real. The following example illustrates the unboundedness of the coefficients of an odd typically-real function:

$$f(z) = \frac{z + z^2}{(1 - z^2)^2} = \frac{1}{2}\left[\frac{z}{(1 - z^2)^2} + \frac{z}{(1 + z)^2}\right] \tag{1.2}$$

$$= z + \sum_{1}^{\infty} (2n + 1) z^{2n+1}.$$

That this function is typically-real is readily seen since the coefficients are real and since the function is also the sum of two univalent (and à fortiori typically-real) functions $\frac{z}{2}(1-z)^{-2}$ and $\frac{z}{2}(1+z)^{-2}$.

In addition, mention might be made here of the fact that the coefficients of an odd typically-real function are bounded if $-if(iz)$ is also typically-real (5).

We are thus led to this question: For the functions $f(z)$ which are typically-real what further property, analogous to $f(z)$ being an odd function, will ensure that the coefficients a_n are uniformly bounded?

The author solves this problem in this paper and finds the best possible value of the universal constant which is the bound for the coefficients.

For the condition

$$f(z)+f(-z)\equiv 0, \quad \text{or} \quad a_n=0 \qquad \text{for } n\equiv 0 \bmod 2 \tag{1.3}$$

which ensured boundedness for the coefficients of univalent functions we substitute the following generalization of (1.3) to be applied to the typically-real functions:

$$\sum_{k=0}^{p-1} f(ze^{2k\pi i/p})\equiv 0, \quad \text{or} \quad a_n=0 \qquad \text{for } n\equiv 0 \bmod p, \tag{1.4}$$

where p is any odd integer greater than one. In this case there exists a constant $A(p)$ independent of $f(z)$ but depending upon p such that for all n

$$|a_n|<A(p). \tag{1.5}$$

Further, (1.5) is true for even p and even n. But it is false for even p and odd n. We find the following estimates for $A(p)$:

Case (a): p odd, n odd, $n<p$,

$$|a_n| \leqq \text{smaller of } (n, p-n) \leqq \frac{p-1}{2}.$$

Case (b): p odd, n odd, $n>p$,

$$|a_n| \leqq \text{smaller of } \left(n-p\left[\frac{n}{p}\right]_o, p\left[\frac{n}{p}\right]_o+2p-n\right) \leqq p-1.$$

Case (c): p odd, n even,

$$|a_n| \leqq \text{smaller of } \left(n-p\left[\frac{n}{p}\right]_e, p\left[\frac{n}{p}\right]_e+2p-n\right) \leqq p-1.$$

Case (d): p even, n even,

$$|a_n| \leqq \text{smaller of } \left(n-p\left[\frac{n}{p}\right], p\left[\frac{n}{p}\right]+p-n\right) \leqq \frac{p}{2}.$$

Thus in every case $A(p) \leqq p - 1$. We have used symbols above which we define as follows:

$$
\begin{aligned}
[x]_o &= \text{the largest odd integer} \leqq x, \\
[x]_e &= \text{the largest even integer} \leqq x, \\
[x] &= \text{the largest integer} \leqq x.
\end{aligned} \tag{1.6}
$$

For the special case $p = 3$ we have the best possible value of $A(p)$. We obtain the sharp result

$$
|a_n| \leqq \underset{0 \leqq \phi \leqq \frac{\pi}{3}}{\text{maximum}} \left| \frac{\cos n\phi}{\cos \phi} \right| < 2. \tag{1.7}
$$

For a fixed n the equality sign is attained by the typically-real function

$$
\begin{aligned}
\Phi(z;\phi) &= \frac{z - z^2 \sec \phi + z^3}{\left\{1 - 2z\cos\left(\frac{\pi}{3} + \phi\right) + z^2\right\}\left\{1 - 2z\cos\left(\frac{\pi}{3} - \phi\right) + z^2\right\}} \\
&= \sum_{n=1}^{\infty} \frac{2}{3^{\frac{1}{2}}} \sin\frac{n\pi}{3} \frac{\cos n\phi}{\cos \phi} z^n
\end{aligned} \tag{1.8}
$$

in which ϕ has the value for which $|\cos n\phi/\cos\phi|$ attains its maximum in the closed interval $(0, \pi/3)$. In the theory of univalent functions the extremal function usually turns out to be the same regardless of the value of n. Hence this example is especially interesting since the function for which $|a_n|$ attains its maximum varies with n. Simple calculations give for $p = 3$:

$$
\begin{aligned}
&|a_2| \leqq 1, \qquad |a_4| \leqq \frac{4}{9} 3^{\frac{1}{2}}(4 - 10^{\frac{1}{2}})(2 + 10^{\frac{1}{2}})^{\frac{1}{2}} = 1.466\cdots, \\
&|a_5| \leqq \frac{5}{4} = 1.25, \qquad a_{3k} = 0, \cdots \\
&\limsup_{n\to\infty} \max_{f(z)} |a_n| = 2,
\end{aligned} \tag{1.9}
$$

where equality signs are attained.

Finally, for the case $p = 3$ we obtain a Stieltjes integral representation for the functions $f(z)$ of the type considered above from which we deduce an inequality for the rate of growth of $f(z)$:

$$
|f(re^{i\theta})| \leqq \frac{4r}{3^{\frac{1}{2}}(1 - r^2)}. \tag{1.10}
$$

2. The case $p = 3$

Let $f(z)$ be given by (1.1) with

$$
a_n = 0 \qquad \text{for } n \equiv 0 \pmod 3.
$$

We suppose $f(z)$ typically-real for $|z| < 1$. Thus the coefficients are real and in addition, if $z = re^{i\theta}$, the imaginary part of $f(z)$, denoted by $\Im f(z)$, has the same sign as $\sin\theta$ for $r < 1$. In particular this condition is satisfied whenever $f(z)$ is univalent and has its coefficients all real. We need the following

LEMMA. *The necessary and sufficient condition that*

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p$$

for the function

$$f(z) = \sum a_n z^n$$

is that the following identity be true:

$$\sum_{k=0}^{p-1} f(ze^{2k\pi i/p}) \equiv 0. \tag{2.1}$$

PROOF. Suppose that (2.1) holds. If $z = re^{i\theta}$ then

$$a_n e^{2nk\pi i/p} = \frac{1}{2\pi r^n}\int_0^{2\pi} f(ze^{2k\pi i/p})e^{-ni\theta}d\theta,$$

$$a_n \cdot \sum_{k=0}^{p-1} e^{2nk\pi i/p} = \frac{1}{2\pi r^n}\int_0^{2\pi} e^{-ni\theta}\cdot\sum_{k=0}^{p-1} f(ze^{2k\pi i/p})\,d\theta,$$

or

$$pa_n = 0 \qquad \text{for } n \equiv 0 \bmod p.$$

Conversely, if

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p,$$

a consideration of the power series of $f(z)$ shows that (2.1) holds in virtue of the relation

$$\sum_{k=0}^{p-1} e^{2mk\pi i/p} = \frac{1 - e^{2m\pi i}}{1 - e^{2m\pi i/p}} = 0, \qquad m \not\equiv 0 \bmod p.$$

This completes the proof of the lemma.

We make use of this lemma at first with $p = 3$. Let

$$\Im f(re^{i\theta}) = v(r, \theta).$$

Since $f(z)$ is typically-real we have for $r < 1$

$$\begin{aligned} &v(r, \theta) > 0 \quad \text{for} \quad 0 < \theta < \pi, \\ &v(r, \theta) < 0 \quad \text{for} \quad \pi < \theta < 2\pi, \\ &v(r, \pi - \theta) = -v(r, \pi + \theta), \\ &v(r, \theta) \qquad = -v(r, -\theta). \end{aligned} \tag{2.2}$$

By (2.1) with $p = 3$ we have

$$f(z) + f(ze^{2\pi i/3}) + f(ze^{4\pi i/3}) \equiv 0,$$

and in particular the imaginary part of the lefthand member is zero. This can be written in the form

$$v(r, \theta) + v\left(r, \frac{2\pi}{3} + \theta\right) - v\left(r, \frac{2\pi}{3} - \theta\right) = 0. \tag{2.3}$$

The coefficients of $f(z)$ are all real and are given by

$$a_n = \frac{2}{\pi r^n} \int_0^\pi v(r, \theta) \sin n\theta \, d\theta. \tag{2.4}$$

We shall modify formula (2.4) by using (2.3), and alter the range of integration. We notice in passing that for $n = 1$ the integrand in (2.4) is positive over the range of integration.

$$\int_0^\pi v(r, \theta) \sin n\theta \, d\theta = \int_0^{\pi/3} + \int_{\pi/3}^{2\pi/3} + \int_{2\pi/3}^{\pi} = I_1 + I_2 + I_3. \tag{2.5}$$

Let $\theta = \frac{2\pi}{3} - \phi$ in I_2 which then becomes

$$I_2 = \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) d\phi. \tag{2.6}$$

Let $\theta = \frac{2\pi}{3} + \phi$ in I_3 which becomes

$$I_3 = \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) d\phi. \tag{2.7}$$

Using (2.6) and (2.7) we write (2.5) as

$$\int_0^{\pi/3} \left\{ v(r, \phi) \sin n\phi + v\left(r, \frac{2\pi}{3} - \phi\right) \sin n\left(\frac{2\pi}{3} - \phi\right) + v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) \right\} d\phi.$$

On substituting the value of $v\left(r, \frac{2\pi}{3} - \phi\right)$ obtained from (2.3) we obtain

$$\int_0^{\pi/3} v(r, \phi) \left\{ \sin n\phi + \sin n\left(\frac{2\pi}{3} - \phi\right) \right\} + v\left(r, \frac{2\pi}{3} + \phi\right) \left\{ \sin n\left(\phi + \frac{2\pi}{3}\right) + \sin n\left(\frac{2\pi}{3} - \phi\right) \right\} d\phi$$

$$= 2 \sin \frac{n\pi}{3} \int_0^{\pi/3} v(r, \phi) \cos n\left(\frac{\pi}{3} - \phi\right) d\phi + 2 \sin \frac{2n\pi}{3} \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \cos n\phi \, d\phi$$

We suppose $f(z)$ typically-real for $|z| < 1$. Thus the coefficients are real and in addition, if $z = re^{i\theta}$, the imaginary part of $f(z)$, denoted by $\Im f(z)$, has the same sign as $\sin\theta$ for $r < 1$. In particular this condition is satisfied whenever $f(z)$ is univalent and has its coefficients all real. We need the following

LEMMA. *The necessary and sufficient condition that*

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p$$

for the function

$$f(z) = \sum a_n z^n$$

is that the following identity be true:

$$\sum_{k=0}^{p-1} f(ze^{2k\pi i/p}) \equiv 0. \tag{2.1}$$

PROOF. Suppose that (2.1) holds. If $z = re^{i\theta}$ then

$$a_n e^{2nk\pi i/p} = \frac{1}{2\pi r^n}\int_0^{2\pi} f(ze^{2k\pi i/p})e^{-ni\theta}d\theta\,,$$

$$a_n \cdot \sum_{k=0}^{p-1} e^{2nk\pi i/p} = \frac{1}{2\pi r^n}\int_0^{2\pi} e^{-ni\theta}\cdot\sum_{k=0}^{p-1} f(ze^{2k\pi i/p})\,d\theta,$$

or

$$pa_n = 0 \qquad \text{for } n \equiv 0 \bmod p.$$

Conversely, if

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p,$$

a consideration of the power series of $f(z)$ shows that (2.1) holds in virtue of the relation

$$\sum_{k=0}^{p-1} e^{2mk\pi i/p} = \frac{1 - e^{2m\pi i}}{1 - e^{2m\pi i/p}} = 0, \qquad m \not\equiv 0 \bmod p.$$

This completes the proof of the lemma.

We make use of this lemma at first with $p = 3$. Let

$$\Im f(re^{i\theta}) = v(r, \theta).$$

Since $f(z)$ is typically-real we have for $r < 1$

$$\begin{aligned} &v(r, \theta) > 0 \quad \text{for} \quad 0 < \theta < \pi, \\ &v(r, \theta) < 0 \quad \text{for} \quad \pi < \theta < 2\pi, \\ &v(r, \pi - \theta) = -v(r, \pi + \theta), \\ &v(r, \theta) \qquad = -v(r, -\theta). \end{aligned} \tag{2.2}$$

By (2.1) with $p = 3$ we have

$$f(z) + f(ze^{2\pi i/3}) + f(ze^{4\pi i/3}) \equiv 0,$$

and in particular the imaginary part of the lefthand member is zero. This can be written in the form

$$v(r, \theta) + v\left(r, \frac{2\pi}{3} + \theta\right) - v\left(r, \frac{2\pi}{3} - \theta\right) = 0. \tag{2.3}$$

The coefficients of $f(z)$ are all real and are given by

$$a_n = \frac{2}{\pi r^n} \int_0^\pi v(r, \theta) \sin n\theta \, d\theta. \tag{2.4}$$

We shall modify formula (2.4) by using (2.3), and alter the range of integration. We notice in passing that for $n = 1$ the integrand in (2.4) is positive over the range of integration.

$$\int_0^\pi v(r, \theta) \sin n\theta \, d\theta = \int_0^{\pi/3} + \int_{\pi/3}^{2\pi/3} + \int_{2\pi/3}^{\pi} = I_1 + I_2 + I_3. \tag{2.5}$$

Let $\theta = \frac{2\pi}{3} - \phi$ in I_2 which then becomes

$$I_2 = \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) d\phi. \tag{2.6}$$

Let $\theta = \frac{2\pi}{3} + \phi$ in I_3 which becomes

$$I_3 = \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) d\phi. \tag{2.7}$$

Using (2.6) and (2.7) we write (2.5) as

$$\int_0^{\pi/3} \left\{ v(r, \phi) \sin n\phi + v\left(r, \frac{2\pi}{3} - \phi\right) \sin n\left(\frac{2\pi}{3} - \phi\right) + v\left(r, \frac{2\pi}{3} + \phi\right) \sin n\left(\frac{2\pi}{3} + \phi\right) \right\} d\phi.$$

On substituting the value of $v\left(r, \frac{2\pi}{3} - \phi\right)$ obtained from (2.3) we obtain

$$\int_0^{\pi/3} v(r, \phi) \left\{ \sin n\phi + \sin n\left(\frac{2\pi}{3} - \phi\right) \right\} + v\left(r, \frac{2\pi}{3} + \phi\right) \left\{ \sin n\left(\phi + \frac{2\pi}{3}\right) + \sin n\left(\frac{2\pi}{3} - \phi\right) \right\} d\phi$$

$$= 2 \sin \frac{n\pi}{3} \int_0^{\pi/3} v(r, \phi) \cos n\left(\frac{\pi}{3} - \phi\right) d\phi + 2 \sin \frac{2n\pi}{3} \int_0^{\pi/3} v\left(r, \frac{2\pi}{3} + \phi\right) \cos n\phi \, d\phi$$

$$= 2\sin\frac{n\pi}{3}\int_0^{\pi/3} v\left(r, \frac{\pi}{3}-\phi\right)\cos n\phi\, d\phi + 2\sin\frac{2n\pi}{3}\int_0^{\pi/3} v\left(r, \frac{2\pi}{3}+\phi\right)\cos n\phi\, d\phi$$

$$= 2\sin\frac{n\pi}{3}\int_0^{\pi/3}\left\{v\left(r, \frac{\pi}{3}-\phi\right) + 2\cos\frac{n\pi}{3}\, v\left(r, \frac{2\pi}{3}+\phi\right)\right\}\cos n\phi\, d\phi.$$

Thus we have modified (2.4) to the new form

$$(2.8)\quad a_n = \frac{4}{\pi r^n}\sin\frac{n\pi}{3}\int_0^{\pi/3}\left\{v\left(r, \frac{\pi}{3}-\phi\right) + 2\cos\frac{n\pi}{3}\, v\left(r, \frac{2\pi}{3}+\phi\right)\right\}\cos n\phi\, d\phi,$$

and in spite of the fact that we have used the identity (2.3), in which there appears a negative term, to obtain (2.8), the integrand in (2.8) for $n = 1$ contains only positive terms over the range of integration, as did (2.4). Therefore, since we also have for $n = 1$

$$1 = a_1 = \frac{2\cdot 3^{\frac{1}{2}}}{\pi r}\int_0^{\pi/3}\left\{v\left(r, \frac{\pi}{3}-\phi\right) + v\left(r, \frac{2\pi}{3}+\phi\right)\right\}\cos\phi\, d\phi$$

we obtain the inequality

$$|a_n| \leqq \frac{2\cdot 3^{\frac{1}{2}}}{\pi r^n}\int_0^{\pi/3}\left|v\left(r, \frac{\pi}{3}-\phi\right) + 2\cos\frac{n\pi}{3}\, v\left(r, \frac{2\pi}{3}+\phi\right)\right|\cdot|\cos n\phi|\, d\phi$$

$$(2.9)\qquad \leqq \frac{2\cdot 3^{\frac{1}{2}}}{\pi r^n}\int_0^{\pi/3}\left\{v\left(r. \frac{\pi}{3}-\phi\right) + v\left(r, \frac{2\pi}{3}+\phi\right)\right\}\left|\frac{\cos n\phi}{\cos\phi}\right|\cos\phi\, d\phi$$

$$\leqq \frac{a_1}{r^{n-1}}\operatorname*{maximum}_{0\leqq\phi\leqq\pi/3}\left|\frac{\cos n\phi}{\cos\phi}\right|.$$

On letting $r \to 1$ we obtain

$$(2.10)\qquad |a_n| \leqq \operatorname*{maximum}_{0\leqq\phi\leqq\pi/3}\left|\frac{\cos n\phi}{\cos\phi}\right|.$$

The inequality (2.10) is sharp inasmuch as the equality sign is attained for fixed n by the function $\Phi(z;\phi)$ defined in (1.8) wherein a proper choice of ϕ is made. To see that $\Phi(z;\phi)$ is typically-real for $|z| < 1$ and for $0 \leqq \phi \leqq \pi/3$ we may write

$$(2.11)\qquad \Im\Phi(re^{i\theta};\phi) = \frac{r}{4}(1-r^2)\sin\theta\cdot\frac{N}{D}$$

where

$$N = \{(1+r^2)\sec\phi - 4r\cos\theta\}^2 + (4-\sec^2\phi)\{(r-\cos 2\phi)^2 + \sin^2 2\phi\}$$

$$D = \left|1 - 2z\cos\left(\frac{\pi}{3}+\phi\right) + z^2\right|^2\cdot\left|1 - 2z\cos\left(\frac{\pi}{3}-\phi\right) + z^2\right|^2.$$

Consequently the imaginary part of $\Phi(re^{i\theta};\phi)$ has the same sign as $\sin\phi$ for $r < 1$ provided $|\sec\phi| \leqq 2$. This last condition is satisfied if ϕ lies in the closed interval $(0, \pi/3)$ which we have used. Although the function $\Phi(z;\phi)$

is typically-real it is, however, univalent only for $\phi = 0$ and for $\phi = \pi/3$ in which cases the function reduces to

$$z(1 - z + z^2)^{-1} \quad \text{and} \quad z(1 + z + z^2)^{-1} \qquad \text{respectively.}$$

Let us now calculate the first few values of μ_n defined as

$$\mu_n = \underset{0 \leqq \phi \leqq \pi/3}{\text{maximum}} \left| \frac{\cos n\phi}{\cos \phi} \right|, \qquad n \not\equiv 0 \bmod 3.$$

We note that local maximum and minimum values of $\dfrac{\cos n\phi}{\cos \phi}$ occur at points ϕ for which

$$\tan \phi = n \tan n\phi, \qquad 0 \leqq \phi \leqq \pi/3, \tag{2.12}$$

at which points we have

$$\left| \frac{\cos n\phi}{\cos \phi} \right| = \left[\frac{1 + \tan^2 \phi}{1 + \left(\dfrac{\tan \phi}{n} \right)^2} \right]^{\frac{1}{2}} \geqq 1. \tag{2.13}$$

If $n = 2$ equation (2.12) has the unique solution $\phi = 0$ in the given range. $|\cos n\phi/\cos \phi|$ has the value one at both ends of the interval. It is also a continuous function of ϕ in this interval. Thus $\mu_2 = 1$ and $|a_2| \leqq 1$. For all larger values of n equation (2.12) has at least one positive solution for ϕ. The number of solutions increases with n. For the continuous function $\cos n\phi/\cos \phi$ vanishes for the values $\phi = \left(\dfrac{2k + 1}{2n} \right) \pi$ where k is a non-negative integer not greater than $\left[\dfrac{2n - 3}{6} \right]$. At the same time this function has either a maximum or minimum point between successive zeros of the function. Since at the maximum and minimum points the expression in (2.13) is not less than one, while at the end points the value of $|\cos n\phi/\cos \phi|$ is one, it follows that the maximum has the value given by (2.13), that is

$$\mu_n = \sec \phi \cdot \left[1 + \left(\frac{\tan \phi}{n} \right)^2 \right]^{-\frac{1}{2}}, \tag{2.14}$$

where ϕ is a certain solution of (2.12). But since when n approaches infinity there are zeros of $\cos n\phi/\cos \phi$, and à fortiori solutions of (2.12), arbitrarily near to $\pi/3$, it follows that

$$\limsup_{n \to \infty} \mu_n = \sec \frac{\pi}{3} = 2. \tag{2.15}$$

Elementary calculations give

$$\begin{aligned} &\mu_2 = 1, \\ &\mu_4 = \tfrac{4}{9} 3^{\frac{1}{2}} (4 - 10^{\frac{1}{2}})(2 + 10^{\frac{1}{2}})^{\frac{1}{2}} = 1.466 \cdots, \\ &\mu_5 = \tfrac{5}{4} = 1.25. \end{aligned} \tag{2.16}$$

3. A Stieltjes integral representation of $f(z)$

Assuming $p = 3$, we have by (2.8) the formula for the coefficients

$$a_n = \frac{4}{\pi\rho^n} \sin\frac{n\pi}{3} \int_0^{\pi/3} \left\{ v\left(\rho, \frac{\pi}{3} - \phi\right) + 2\cos\frac{n\pi}{3}\, v\left(\rho, \frac{2\pi}{3} + \phi\right)\right\} \cos n\phi\, d\phi$$

for $\rho < 1$. Hence for $|z| = r < \rho$ we have

$$\begin{aligned} f\left(\frac{z}{\rho}\right) &= \frac{2}{\pi}\int_0^{\pi/3} v\left(\rho, \frac{\pi}{3} - \phi\right)\left\{\sum_{n=1}^{\infty} 2\sin\frac{n\pi}{3}\cos n\phi \left(\frac{z}{\rho}\right)^n\right\} d\phi \\ &\quad + \frac{2}{\pi}\int_0^{\pi/3} v\left(\rho, \frac{2\pi}{3} + \phi\right)\left\{\sum_{n=1}^{\infty} 2\sin\frac{2n\pi}{3}\cos n\phi \left(\frac{z}{\rho}\right)^n\right\} d\phi \\ &= \frac{2\cdot 3^{\frac{1}{2}}}{\pi}\int_0^{\pi/3} \Phi\left(\frac{z}{\rho}; \phi\right)\cdot v\left(\rho, \frac{\pi}{3} - \phi\right)\cos\phi\, d\phi \\ &\quad - \frac{2\cdot 3^{\frac{1}{2}}}{\pi}\int_0^{\pi/3} \Phi\left(\frac{-z}{\rho}; \phi\right)\cdot v\left(\rho, \frac{2\pi}{3} + \phi\right)\cos\phi\, d\phi, \end{aligned}$$

where

$$\Phi(z;\phi) \equiv \frac{z - z^2\sec\phi + z^3}{\left\{1 - 2z\cos\left(\frac{\pi}{3} + \phi\right) + z^2\right\}\left\{1 - 2z\cos\left(\frac{\pi}{3} - \phi\right) + z^2\right\}}. \tag{3.1}$$

We define two functions of bounded variation by the equations

$$\begin{aligned} \alpha(\rho, \phi) &= \frac{2}{3^{\frac{1}{2}}}\int_0^{\phi} v\left(\rho, \frac{\pi}{3} - \phi\right)\cos\phi\, d\phi, \\ \beta(\rho, \phi) &= \frac{2}{3^{\frac{1}{2}}}\int_0^{\phi} v\left(\rho, \frac{2\pi}{3} + \phi\right)\cos\phi\, d\phi, \end{aligned} \tag{3.2}$$

defined for the range $0 \leqq \phi \leqq \pi/3$. Since the integrands are positive these functions are non-decreasing functions of ϕ for fixed $\rho < 1$. As ρ varies the functions $\alpha(\rho, \phi)$ of ϕ are of bounded variation uniformly in ρ since the total variation is

$$\begin{aligned} \int_0^{\pi/3} d\alpha(\rho, \phi) &= \frac{2}{3^{\frac{1}{2}}}\int_0^{\pi/3} v\left(\rho, \frac{\pi}{3} - \phi\right)\cos\phi\, d\phi \\ &< \frac{2}{3^{\frac{1}{2}}}\int_0^{\pi/3} \left\{v\left(\rho, \frac{\pi}{3} - \phi\right) + v\left(\rho, \frac{2\pi}{3} + \phi\right)\right\}\cos\phi\, d\phi \\ &\leqq \pi\rho a_1/3 < \pi/3. \end{aligned}$$

A similar remark applied to $\beta(\rho, \phi)$. Consequently for any sequence $\{\rho_m\}$ approaching unity through values less than one the functions

$$\alpha_m(\phi) \equiv \alpha(\rho_m, \phi)$$

are uniformly of bounded variation. It follows from a theorem of Helly (6) that

the $\alpha_m(\phi)$ are a compact set: there exists a subsequence $\alpha_n(\phi)$ of the sequence $\alpha_m(\phi)$ and a limiting function $\alpha(\phi)$ (in our case non-decreasing also) such that

$$\lim_{n\to\infty} \alpha_n(\phi) = \alpha(\phi)$$

for every ϕ in the closed interval $(0, \pi/3)$.

By another theorem of Helly (6) it then follows that

$$f(z) = \lim_{\rho_n\to 1} f(z\rho_n)$$

$$= \lim_{\rho_n\to 1} \frac{3}{\pi}\int_0^{\pi/3} \{\Phi(z;\phi)d\alpha(\rho_n,\phi) - \Phi(-z;\phi)d\beta(\rho_n,\phi)\},$$

$$f(z) = \frac{3}{\pi}\int_0^{\pi/3} \{\Phi(z;\phi)d\alpha(\phi) - \Phi(-z;\phi)d\beta(\phi)\}, \tag{3.3}$$

where $\alpha(\phi)$ and $\beta(\phi)$ are non-decreasing functions of ϕ in the interval $(0, \pi/3)$.

Conversely, if $\alpha(\phi)$ and $\beta(\phi)$ are any two arbitrary non-decreasing functions of ϕ in the closed interval $(0, \pi/3)$, normalized so that

$$\frac{3}{\pi}\int_0^{\pi/3} d\{\alpha(\phi) + \beta(\phi)\} = 1, \tag{3.4}$$

then the function $f(z)$ formed as in (3.3) is a member of the class of functions under discussion. This follows readily since $\Phi(z;\phi)$ and $-\Phi(-z;\phi)$ are both typically-real for $|z| < 1$, and since each has its leading coefficient positive (same sign). Thus formula (3.3) generates all of the functions of our class.

From (3.3) we may now express the coefficients a_n of $f(z)$ by the formula

$$a_n = \frac{\sin\frac{n\pi}{3}}{\sin\frac{\pi}{3}}\cdot\frac{3}{\pi}\int_0^{\pi/3}\frac{\cos n\phi}{\cos\phi}\,d\{\alpha(\phi) + (-1)^{n-1}\beta(\phi)\}. \tag{3.5}$$

From this formula and (3.4) the inequality (1.7) may easily be obtained again.

4. The rate of growth of $f(z)$

From (3.3) and (3.4) we obtain the following inequality for $f(z)$:

$$|f(re^{i\theta})| \leqq \max_{0\leqq\phi\leqq\pi/3}\ \max_{0\leqq\theta<2\pi} |\Phi(re^{i\theta},\phi)| \tag{4.1}$$

where $\Phi(z;\phi)$ is given by (3.1). To calculate an inequality for the right-hand member of (4.1) we write $\Phi(z;\phi)$ as follows with $z = re^{i\theta}$:

$$\frac{2\cdot 3^{\frac{1}{2}}\Phi(z;\phi)}{z} = \frac{3^{\frac{1}{2}} + \tan\phi}{1 - 2z\cos\left(\frac{\pi}{3}+\phi\right) + z^2} + \frac{3^{\frac{1}{2}} - \tan\phi}{1 - 2z\cos\left(\frac{\pi}{3}-\phi\right) + z^2},$$

$$\frac{2\cdot 3^{\frac{1}{2}}}{r}\cdot|\Phi(re^{i\theta};\phi)| \leqq \frac{3^{\frac{1}{2}} + \tan\phi}{\left|1 - 2z\cos\left(\frac{\pi}{3}+\phi\right) + z^2\right|} + \frac{3^{\frac{1}{2}} - \tan\phi}{\left|1 - 2z\cos\left(\frac{\pi}{3}-\phi\right) + z^2\right|},$$

$$|1 - 2z\cos\alpha + z^2| = [(2r\cos\theta - \cos\alpha - r^2\cos\alpha)^2 + (1 - r^2)^2\sin^2\alpha]^{\frac{1}{2}}$$
$$\geqq (1 - r^2)\,|\sin\alpha|,$$

$$\frac{2\cdot 3^{\frac{1}{2}}}{r}\,|\Phi(re^{i\theta};\phi)| \leqq (1 - r^2)^{-1}\left[\frac{3^{\frac{1}{2}} + \tan\phi}{\sin\left(\frac{\pi}{3} + \phi\right)} + \frac{3^{\frac{1}{2}} - \tan\phi}{\sin\left(\frac{\pi}{3} - \phi\right)}\right] = \frac{4\sec\phi}{1 - r^2}.$$

Hence, since $0 \leqq \phi \leqq \pi/3$

$$|\Phi(re^{i\theta};\phi)| \leqq \frac{2}{3^{\frac{1}{2}}}\cdot\frac{r\sec\phi}{1 - r^2} \leqq \frac{4r}{3^{\frac{1}{2}}(1 - r^2)},$$

(4.2)

$$|f(re^{i\theta})| \leqq \frac{4r}{3^{\frac{1}{2}}(1 - r^2)}.$$

The results of the preceding sections apply à fortiori to univalent functions which are real on the real axis and for which (1.4) holds ($p = 3$) inasmuch as these functions are also typically-real.

Again, if we have a univalent function

$$F(z) = z + \sum_1^\infty c_n z^n$$

with real coefficients c_n satisfying

$$c_n = 0 \qquad \text{for } n \equiv 0 \bmod 3,$$

and such that $F(z)$ is convex in the direction of the imaginary axis, certain analogous results can be obtained for $F(z)$. This follows from the relation (7) of $F(z)$ to $f(z)$ given by

$$F(z) = \int_0^z \frac{f(z)}{z}\,dz = \sum_1^\infty \frac{a_n}{n} z^n.$$

Thus $c_n = \dfrac{a_n}{n} = 0\left(\dfrac{1}{n}\right)$ from (1.7).

5. Generalization to $p > 3$

Let $f(z)$ be typically-real in the unit circle, given by (1.1), and let

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p$$

where p is some positive integer greater than one. The question arises whether there is an inequality for a_n analogous to (1.7). To begin with, we may notice that no such inequality implying uniform boundedness to a_n for all n exists for any *even* value of the interger p. This follows at once since the function given by (1.2) satisfies the condition

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod 2k$$

for any positive integer k, yet the coefficients for this function are unbounded. In what follows we therefore restrict ourselves, unless we state otherwise, to odd values of p.

A proof analogous to that given in (2.11) for $\Phi(z;\phi)$ shows that the function

$$\Phi_p(z) \equiv \frac{z - 2z^2\cos\frac{\pi}{p}\sec\phi + z^3}{\left\{1 - 2z\cos\left(\frac{\pi}{p}+\phi\right)+z^2\right\}\left\{1 - 2z\cos\left(\frac{\pi}{p}-\phi\right)+z^2\right\}}$$

(5.1)

$$= \sum_{n=1}^{\infty} \frac{\sin\frac{n\pi}{p}}{\sin\frac{\pi}{p}}\cdot\frac{\cos n\phi}{\cos\phi}\cdot z^{n},$$

is typically-real for $|z| < 1$ for all values of ϕ in the interval $(0, \pi/p)$. This suggests the inequality

$$|a_n| \leqq \frac{\left|\sin\frac{n\pi}{p}\right|}{\sin\frac{\pi}{p}}\cdot \max_{0\leq\phi\leq\pi/p}\left|\frac{\cos n\phi}{\cos\phi}\right| \tag{5.2}$$

as the generalization of (1.7) to $p > 3$. The inequality (5.2) is correct for $p = 2, 3$ and ∞, as is readily seen, with equality sign for the function $\Phi_p(z)$ of (5.1) with an appropriate choice of ϕ. However, we have shown that the inequality (5.2) cannot be correct for even values of p greater than 2.

Whether (5.2) is correct for odd values of $p > 3$ the writer is unable to say. The method employed for $p = 3$ can be generalized to $p > 3$ and leads to the following formula for a_n :

$$a_n = \frac{4}{\pi r^n}\sin\frac{n\pi}{p}\int_0^{\pi/p}\left\{v\left(r,\frac{\pi}{p}-\phi\right)+2\cos\frac{n\pi}{p}\,v\left(r,\frac{2\pi}{p}+\phi\right)\right\}\cos n\phi\,d\phi$$

$$+\frac{4}{\pi r^n}\int_0^{\pi/p}\sum_{k=2}^{\frac{1}{2}(p-1)}\left[\begin{array}{l} v\left(r,\frac{2k\pi}{p}+\phi\right)\sin n(k+1)\frac{\pi}{p}\cos n\left\{(k-1)\frac{\pi}{p}-\phi\right\} \\ +v\left(r,\frac{2k\pi}{p}+\phi\right)\sin n(k-1)\frac{\pi}{p}\cos n\left\{(k+1)\frac{\pi}{p}-\phi\right\}\end{array}\right]d\phi, \tag{5.3}$$

when use is made of (1.4) which may be written in the form

$$v(r,\phi)+\sum_{k=1}^{\frac{1}{2}(p-1)}v\left(r,\frac{2k\pi}{p}+\phi\right)-\sum_{k=1}^{\frac{1}{2}(p-1)}v\left(r,\frac{2k\pi}{p}-\phi\right)\equiv 0. \tag{5.4}$$

The method, however, fails for $p > 3$ since the proof for $p = 3$ necessitated that the integrand in (2.8) be positive when $n = 1$. The integrand in (2.4) was positive for $n = 1$ to begin with, and even after the substitution from (2.3), which contained but one negative term, the integrand in (2.8) remained positive for $n = 1$. But for $p > 3$ the analogous identity (5.4) has more than one negative term and thus the terms of the integrand in (5.3) fail to remain all positive after substitution from (5.4).

That the coefficients are bounded is, however, true for all odd p. The follow-

ing elementary proof yields (1.5), though perhaps not the best possible value of $A(p)$.

By a result due to W. Rogosinski (4)

$$f(z) = \frac{zF(z)}{1 - z^2}, \tag{5.5}$$

where

$$F(z) = 1 + \sum_1^\infty b_n z^n, \qquad \Re F(z) > 0 \qquad \text{for } |z| < 1, \tag{5.6}$$

and where

$$|b_n| \leqq 2 \qquad \text{for all } n. \tag{5.7}$$

From (5.5) we have

$$a_n = 1 + b_2 + b_4 + \cdots + b_{n-1}, \qquad n \text{ odd},$$

$$a_n = b_1 + b_3 + b_5 + \cdots + b_{n-1}, \qquad n \text{ even}.$$

Let us suppose p odd, n odd. Since

$$a_n = 0 \qquad \text{for } n \equiv 0 \bmod p$$

we have

$$1 + b_2 + b_4 + \cdots + b_{p-1} = a_p = 0,$$

$$b_{(2k-1)p+1} + b_{(2k-1)p+3} + \cdots + b_{(2k+1)p-1} = a_{(2k+1)p} - a_{(2k-1)p} = 0, \tag{5.8}$$

for $k = 1, 2, 3, \cdots$, and in addition,

$$b_1 + b_3 + b_5 + \cdots + b_{2p-1} = a_{2p} = 0,$$

$$b_{2kp+1} + b_{2kp+3} + \cdots + b_{(2k+2)p+1} = a_{(2k+2)p} - a_{2kp} = 0. \tag{5.9}$$

In the first place, if n is odd and $n < p$

$$a_n = 1 + b_2 + b_4 + \cdots + b_{n-1},$$

and again

$$a_n = (1 + b_2 + b_4 + \cdots + b_{p-1}) - (b_{n+1} + b_{n+3} + \cdots + b_{p-1})$$

$$= -(b_{n+1} + b_{n+3} + \cdots + b_{p-1}).$$

Thus a_n can be represented as the sum of either $\frac{1}{2}(n + 1)$ or $\frac{1}{2}(p - n)$ terms each of absolute value not exceeding 2 (in the case of $b_0 = 1$, not exceeding one). Consequently,

$$|a_n| \leqq \text{smaller of } (n, p - n) \leqq \frac{p - 1}{2}, \qquad n < p. \tag{5.10}$$

Again, let us suppose n odd, p odd, and $n > p$. Let us define the symbols $[x]_0$, $[x]_e$, $[x]$ as in (1.6). Using (5.8) we have

(5.11) $$a_n = b_{(2k-1)p+1} + \cdots + b_{n-1}$$

for a certain integer k such that $(2k-1)p + 1 \leqq n - 1 \leqq (2k+1)p - 3$ which is to say $2k - 1 = \left[\frac{n}{p}\right]_0$. Again we may alternatively write for a_n :

$$a_n = (b_{(2k-1)p+1} + \cdots + b_{(2k+1)p-1}) - (b_{n+1} + \cdots + b_{(2k+1)p-1})$$
$$= -(b_{n+1} + \cdots + b_{(2k+1)p-1}).$$

Thus a_n may be represented as the sum of either

$$\frac{n - p\left[\frac{n}{p}\right]_0}{2} \quad \text{or} \quad \frac{p\left[\frac{n}{p}\right]_0 + 2p - n}{2}$$

terms of absolute value not exceeding 2. Hence for $n > p$, both n and p odd, we have

(5.12) $$|a_n| \leqq \text{smaller of} \left(n - p\left[\frac{n}{p}\right]_0, p\left[\frac{n}{p}\right]_0 + 2p - n\right) \leqq p - 1.$$

Next, let us suppose p odd, n even.

$$a_n = b_1 + b_3 \cdots + b_{n-1} = b_{\nu+1} + \cdots + b_{n-1}, \qquad \nu = p\left[\frac{n}{p}\right]_e$$

$$a_n = -(b_{n+1} + \cdots + b_{\nu+2p-1})$$

whence

(5.13) $$|a_n| \leqq \text{smaller of} \left(n - p\left[\frac{n}{p}\right]_e, p\left[\frac{n}{p}\right]_e + 2p - n\right) \leqq p - 1.$$

Thus for p odd, we have for all n

(5.14) $$|a_n| \leqq p - 1.$$

It is clear by section two that for $p = 3$ the inequality (5.14) cannot be replaced by

$$|a_n| \leqq p - 1 - \epsilon$$

for $\epsilon > 0$ arbitrarily small.

Although for even values of p the coefficients are unbounded, nevertheless, the coefficients with *even* suffixes are bounded. To see this assume p even and proceed as follows:

$$b_{kp+1} + b_{kp+3} + \cdots + b_{(k+1)p+1} = a_{(k+1)p} - a_{kp} = 0$$

for $k = 1, 2, 3, \cdots$. Hence, if n and p are even,

$$a_n = b_{\mu+1} + \cdots + b_{n-1} = -(b_{n+1} + \cdots + b_{\mu+p-1})$$

where

$$\mu = p\left[\frac{n}{p}\right].$$

$$|a_n| \leqq \text{smaller of } \left(n - p\left[\frac{n}{p}\right], p\left[\frac{n}{p}\right] + p - n\right) \leqq \frac{p}{2}. \tag{5.15}$$

A consideration of the several inequalities above yields the following

THEOREM. *If*

$$f_n(z) = z + a_2z^2 + a_3z^3 + \cdots + a_nz^n$$

is a polynomial of degree n and is typically-real with respect to the unit circle the coefficients satisfy the inequality:

$$|a_k| \leqq \left[\frac{n+2}{2}\right], \qquad k = 1, 2, \cdots n.$$

This completes the proof of the statements in the introduction to this paper.

RUTGERS UNIVERSITY,
NEW BRUNSWICK, N. J.

REFERENCES

(1) J. E. LITTLEWOOD, *On inequalities in the theory of functions.* Proc. Lond. Math. Soc., (2), vol. 23, (1925), pp. 481–519.

(2) R. E. A. C. PALEY AND J. E. LITTLEWOOD, *A proof that an odd schlicht function has bounded coefficients*, Jour. Lond. Math. Soc., vol. 7, (1932), pp. 167–169.

(3) J. DIEUDONNÉ, *Polynomes et fonctions bornées d'une variable complexe*, Annales de l'École Normale, vol. 48 (1931), pp. 247–358.

(4) W. ROGOSINSKI. *Über positive harmonische Entwicklungen und typisch-reelle Potenzreihen*, Math. Zeits., vol. 35 (1932), pp. 93–121.

(5) M. S. ROBERTSON. *On the coefficients of a typically-real function*, Bull. Am. Math. Soc., vol. 41 (1935), pp. 565–572.

(6) E. HELLY. *Über lineare Funktionaloperationen*, Wiener Sitzungsberichte, vol. 121 (1912), pp. 265–297.

(7) M. S. ROBERTSON, *On the theory of univalent functions*, Annals of Math., vol. 37 (1936), pp. 374–408.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

A GENERALIZED LAMBERT SERIES AND ITS MOEBIUS FUNCTION[1]

BY WILLIAM C. DOYLE, S.J.

(Received November 23, 1938)

I. Introduction

The series which we shall discuss is

$$H(z) = \sum \frac{a_n b_n z^{\lambda n}}{1 - a_n z^{\mu n}},$$ [2]

where λ is any integer, positive, negative, or zero; μ, a positive integer, greater than zero; and a_n and b_n, any two sequences of complex numbers. This function, which we shall call the H-series, is a generalization of Garvin's F-series[3]

$$\sum \frac{a_n z^{\lambda n}}{1 - z^{\mu n}}$$

and Feld's series[4]

$$\sum \frac{a_n b_n z^n}{1 - a_n z^n};$$

consequently, as pointed out by Garvin, all other Lambert series, such as Weierstrass' and Hansen's, are particular cases of the H-series. It is possible, of course, to define $H(z)$ for any integral value of μ, but this would needlessly complicate the statement of most of the following theorems. Instead, a special theorem is given for negative μ's. When μ is zero, the H-series reduces to a simple power series.

After determining the region of ordinary and uniform convergence of the H-series, its relation to power series will be discussed and a new generalized Moebius function developed to make possible the expansion of any power series in an H-series.

II. The Region of Convergence

In a preliminary study it is natural to restrict ourselves to the regions where $H(z)$ is certainly regular. For this purpose let

[1] This study was made under the guidance of Francis Regan of St. Louis University, St. Louis, Mo., and presented to the American Mathematical Society in two papers under the titles, "On Garvin's $R(n)$ function," (April 10, 1937) and "A further generalization of Lambert series" (September 7, 1937).

[2] Unless otherwise specified, all summations are from $n = 1$ to $n = \infty$.

[3] M. C. Garvin, "A Generalized Lambert Series," American Jour. of Math., LVIII, 3 (1936), pp. 507–513.

[4] J. M. Feld, "The Expansion of Analytic Functions in Generalized Lambert Series," Annals of Math., XXXIII (1932), p. 139.

$$L \leqq |a_n|^{1/n} \leqq M,$$

then the region or ring in which the singularities are, is

$$1/M^{1/\mu} \leqq |z| \leqq 1/L^{1/\mu}.$$

At first thought it might appear that the exclusion of this ring is significant; but this is not so. In fact, even the unbounded sequence, $a_n = n$, makes the series have only a small region of singularities, namely $(1/3)^{1/3\mu} \leqq |z| \leqq 1$.

Note that this singularity ring degenerates into the unit circle of the F-series when $a_n = 1$, or into some other circle when $a_n = a$. Furthermore if L is zero, the region covers all of the complex plane except that portion which is included in a circle of radius, $1/M^{1/\mu}$, with center at the origin.

By means of the test of du Bois-Reymond,[5] the regions of ordinary convergence are found to be as follows,

THEOREM I. *The H-series converges and diverges at the same points as* $\sum a_n b_n z^{\lambda n}$ *for* $|z| < 1/M^{1/\mu}$.

THEOREM II. *The H-series converges and diverges at the same points as* $\sum b_n z^{(\lambda-\mu)n}$ *for* $|z| > 1/L^{1/\mu}$.

THEOREM III. *If* μ *is a negative integer, the H-series converges and diverges at the same points as* $\sum a_n b_n z^{\lambda n}$ *for* $|z| > 1/M^{1/\mu}$ *and at the same points as* $\sum b_n z^{(\lambda-\mu)n}$ *for* $|z| < 1/L^{1/\mu}$.

The third theorem follows immediately from the two preceding if the transformation $z = 1/z'$ is applied to the H-series. We will merely indicate the proof of Theorem II, and leave it to the reader to supply the details of the method to Theorem I.

Since

$$\sum b_n z^{(\lambda-\mu)n} = \sum \frac{a_n b_n z^{\lambda n}}{1 - a_n z^{\mu n}} \cdot \frac{1 - a_n z^{\mu n}}{a_n z^{\mu n}} = \sum A_n B_n,$$

we know from the du Bois-Reymond test that $\sum b_n z^{(\lambda-\mu)n}$ converges if A_n (i.e. $H(z)$) converges and $\sum (B_{n+1} - B_n)$ converges absolutely. But

$$\sum |B_{n+1} - B_n| < \sum 1/(Lz^{\mu})^n \cdot (|z|^{-\mu} + 1)$$

does converge when $|Lz^{\mu}| < 1$. In a similar manner we may also show that $H(z)$ converges if $\sum b_n z^{(\lambda-\mu)n}$ converges. For this purpose write

$$H(z) = \sum b_n z^{(\lambda-\mu)n} \left(\frac{1}{1 - a_n z^{\mu n}} - 1 \right).$$

It is now an easy matter to apply the Weierstrass M-test and prove

THEOREM IV. *The H-series converges uniformly in every closed region lying completely within one of its regions of convergence.*

[5] K. Knopp, "Theory and Application of Infinite Series," p. 430.

III. Some Transformations

In order to facilitate both the application of the above theorems and further theoretical study of the series itself, it is very useful to have at hand the following transformations,

$$(1) \qquad z = s/M^{1/\mu} \qquad z = s/L^{1/\mu} \qquad z = s/(ML)^{1/2\mu}.$$

The result of these substitutions is a new H-series with M or L equal to one or, in the last case, $M = 1/L$. But of prime importance is

$$(2) \qquad zs = (1/(ML))^{1/\mu},$$

an inversion across the circle $|Z| = (1/(ML))^{1/2\mu}$ followed by reflection. The resulting series is

$$(3) \qquad H_1(s) = \sum \frac{b_n(ML)^{(\mu-\lambda)n/\mu} s^{(\mu-\lambda)n}}{1 - \dfrac{(ML)^n}{a_n} s^{\mu n}}.$$

This H-series has exactly the same singularity ring as the original series, but any region of convergence that was beyond the ring of singularities is now within it and conversely.

The last result suggests the possibility of showing that any statement made about an H-series for the region including the origin can be immediately extended to any other possible region. But this is not quite true. When μ is less than λ, relation (3) gives an H-series that diverges at the origin; and this difficulty appears in fact for any H-series that has a negative λ. To avoid this we can use

THEOREM V. *Any H-series with a region of convergence that does not include the origin can be expressed as the sum of another H-series that does include the origin plus a finite number of power series with negative coefficients.*

In proof of this statement we have only to consider the case of an H-series with negative λ's, since any other types can be reduced immediately to (3) (which includes the origin except when $\mu - \lambda$ is negative). The desired expression is given by a generalization of a lemma due to Ananda-Rau,[6] which is

$$(4) \qquad \sum \frac{a_n b_n z^{-\lambda n}}{1 - a_n z^{\mu n}} = \sum \frac{a_n b_n z^{(q\mu-\lambda)n}}{1 - a_n z^{\mu n}} + \sum_{p=1}^{q-1} \sum_{n=1}^{\infty} a_n b_n z^{(p\mu-\lambda)n}, \qquad \lambda > 0.$$

What we wish to show here is that if the left term of (4) has a region within the singularity ring, then a choice of q such that $q\mu - \lambda$ is positive will make the H-series on the right convergent at the origin, and the common region of all the series on the right will be the same as the region of the original series on the left.

[6] K. Ananda-Rau, "On Lambert's Series," Proc. London Math. Soc., ser 2, vol. 18 (1920), p. 3.

For simplicity let us assume that $M = 1$ (use transformation (1) if it is not). Then by Theorem I $\sum a_n b_n z^{-\lambda n}$ converges for $|z| > r$; and since $r < 1$ (i.e. the series in question has by supposition a region within the singularity ring), it is true that $\sum a_n b_n$ converges. Consequently $\sum a_n b_n z^{(q\mu-\lambda)n}$ converges for $|z| < 1$, and the H-series on the right side of (4) converges everywhere within the region of singularities. Finally an application of Cauchy's root test will show that the common region of convergence of the remaining power series is $|z| > r$.

Hence all further considerations of the H-series can be limited to the region including the origin, and extended by the above transformation to the other cases.

III. Relation to Power Series

Since each term of an H-series with a positive λ is analytic for $|z| < r$ and the series itself converges uniformly for $|z| \leqq \rho < r$, it is possible to expand it in a power series. The result is

THEOREM VI. *Every H-series that converges within a circle about the origin can be developed in the series*

$$H(z) = \sum A_n z^n$$

where only those A_n occur whose subscript satisfy the congruence

$$n/d \equiv \lambda \text{ (modulo } u). \tag{5}$$

The values of the A_n are given by the sum

$$A_n = \sum_d{}' a_d^{l+1} b_d \tag{6}$$

over all d's satisfying (5) *and*

$$l = \frac{n - d\lambda}{d\mu}. \tag{7}$$

Next we can ask, under what conditions is formula (6) reversible, that is, given a power series, is it possible to develop it in an H-series? This question is answered by

THEOREM VII. *The necessary and sufficient condition that a power series $\sum A_n z^n$ be expressible as an H-series is that μ be a multiple of λ.*

The proof of this theorem is an exact parallel of Garvin's proof for the corresponding theorem of the F-series.[7] It consists in showing that if μ and λ are relatively prime or are multiples of relatively prime integers, or if λ is a multiple of μ, then the $a_n b_n$ of the H-series is not in general uniquely expressible in terms of the A_n's of its power series. But if $\mu = m\lambda$, $m = 1, 2, 3, \cdots$, then

$$A_{\lambda n} = \sum_d{}' a_d^{l+1} b_d, \tag{8}$$

[7] Am. Jour. of Math., LVIII, 3 (1936) p. 510 (Theorem 3).

where $\sum_d'$ means that the sum is taken over all divisors, d, of n such that $n/d \equiv 1$ (modulo m). This formula is reversible.

We preface the development of the expression for this reversibility with a definition of a generalized Moebius coefficient.

IV. The Number Theory of m-Divisors

Formula (8) is characterized by the fact that the sum is taken over those divisors, d, of n that make $n/d \equiv 1$ (modulo m). It is easy to show that if $n \equiv 1$ (modulo m) these divisors always belong to the class of numbers congruent to one modulo m. For this reason we shall use the letters u, d, and h for these numbers and make the following

DEFINITION. *The m-divisors of u are those divisors of u that are congruent to one modulo m.*

Two interesting and easily proved properties of these divisors are

Both d and u/d are m-divisors of u.

If d is an m-divisor of u, and h is an m-divisor of d, then h is an m-divisor of u.

The Moebius coefficient for m-divisors is Garvin's $R(n)$ function. This can be generalized by what we shall call the S-function modulo m, defining $S(u)$ as a *function of a* such that

$$\sum_d' a^{l(u/d)} S(d) = \begin{cases} 1 \text{ if } u = 1 \\ 0 \text{ if } u > 1 \end{cases}.$$

For example if $m = 3$, the values of $S(u)$ can be built up in the same manner as $R(n)$, and are found to be

$$S(1) = 1 \qquad S(16) = a^2 - a^5$$
$$S(4) = -a \qquad \text{etc.}$$
$$S(7) = -a^2$$
$$S(10) = -a^3 \qquad S(28) = 2a^3 - a^9$$
$$S(13) = -a^4 \qquad S(112) = -a^{37} + 2a^{10} + 2a^7 - 3a^4.$$

It is also possible to evaluate $S(u)$ directly for any u by means of the formula

$$S(u) = \sum_r (-1)^r \sum_P a^{(\Sigma' d - r)/m}, \tag{9}$$

where d means here the set of m-divisors of u greater than one; $\sum' d$, the sum of the r d's whose product is u; the sum, $\sum_r$, is taken over all integral values of r that have any meaning; and the sum, $\sum_P$, taken over each permutation of $\sum' d$.

To verify this formula take the above definition of $S(u)$ for a general value of u, denoting the m-divisors by $1, d_1, d_2, d_3, \cdots$; then insert the expression of each $S(d_i)$, expanded in the same manner. Solving this expression for $S(u)$ and multiplying out the parentheses, we obtain

$$S(u) = -a^{l(u)} + a^{l(u/d_1)+l(d_1)} + a^{l(u/d_1)+l(d_1/d_{11})}S(d_{11}) + a^{l(u/d_1)+l(d_1/d_{12})}S(d_{12})$$
$$+ \cdots + a^{l(u/d_2)+l(d_2)} + a^{l(u/d_2)+l(d_2/d_{21})}S(d_{21}) + a^{l(u/d_2)+l(d_2/d_{22})}S(d_{22})$$
$$+ \cdots + \cdots + \cdots$$

where d_{i1}, d_{i2}, d_{i3}, $\cdots$ are the m-divisors of d_i. Continue this process and regroup the terms according to the number of l's in the exponent of a. When $l(d)$ is replaced by $(d-1)/m$, it is an easy matter to derive formula (9).

This S-function modulo m will be used as the inversion relation for the H-series when $a_n = a$; but in the general case it will be necessary to use a wider definition of $S(u)$, namely,

$$\sum_d{}' a_{n/u}^{l(u/d)} S(d) = \begin{cases} 1 \text{ if } u = 1 \\ 0 \text{ if } u > 1 \end{cases}. \tag{10}$$

In this case each $S(u)$ must be built up from the definition.

When $m = 1$ and $a_n = 1$, the S-function reduces immediately to the Moebius coefficient. It is Garvin's R-function if $a_n = 1$ and $m \neq 1$. In the latter case formula (9) can be used to good advantage in making a new definition of $R(n)$ that has some resemblance to the definition of the original Moebius function.

V. The Expression of a Power Series as an H-Series

We are now in a position to develop the general inversion theorem for m-divisors expressed in terms of the coefficients of the H-series and a corresponding power series.

Formula (8) can be written in this form

$$A_{\lambda d} = \sum_h{}' a_h^{l(d/h)+1} b_h .$$

Multiplying both members of this equation by $S(n/d)$ and summing over all m-divisors of n, we obtain

$$\sum_d{}' S(n/d) A_{\lambda d} = \sum_d{}' S(n/d) \sum_h{}' a_h a_h^{l(d/h)} b_h .$$

Now in the double sum of the right hand member of this equation there is a set of d's that are the m-divisors of n, and for each d, a set of h's that are m-divisors of d. Consider for a moment the complete set of h's, namely, the h's corresponding to all the d's. On the one hand there are as many h's as there are d's, since each d is an m-divisor of d, and is therefore an h. On the other hand there are no more, since each h is also a d. It is true then that the complete set of h's is the same as the set of d's.

It is possible, then to rearrange the above equation in the following manner

$$\sum_d{}' S(n/d) A_{\lambda d} = \sum_d{}' a_d b_d \sum_h{}' S(n/(hd)) a_h^{l(h)} .$$

Since by the definition of the S-function every term of the right hand side is zero except when d is equal to n, the whole expression reduces to $a_n b_n$, and we have proved

THEOREM VIII. *The coefficient $a_n b_n$ of the H-series that corresponds to the series $\sum A_{\lambda n} z^{\lambda n}$ is given by the equation,*

$$a_n b_n = \sum_d' A_{\lambda d} S(n/d).$$

ROCKHURST COLLEGE, KANSAS CITY, MISSOURI.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

A TYPE OF ALGEBRAIC CLOSURE

BY MARSHALL HALL

(Received July 21, 1938)

1. Introduction

In a previous paper[1] the author was led to the following question: Given the complete set of solutions $\{X\} = [x_1, x_2, \cdots, x_n]$ of linear equations

$$L_i = a_1^{(i)}x_1 + \cdots + a_n^{(i)}x_n = 0, \qquad i = 1, \cdots, s, \tag{1.1}$$

over an algebra $\mathfrak{A}$, to what extent can we determine the equations? Naturally the solutions $\{X\}$ will also satisfy any equation $\sum\lambda_i L_i = 0$ derivable from (1.1) with any λ_i from $\mathfrak{A}$. Write $A_i = [a_1^{(i)}, \cdots, a_n^{(i)}]$. Then $S_l = \sum\lambda_i A_i$, $\lambda_i \subset \mathfrak{A}$ is a left vector ideal over $\mathfrak{A}$ and the solutions $\{X\}$ belong to all linear equations whose coefficients are vectors of S_l. We shall express equations (1.1) by saying that the vectors $\{X\}$ annihilate the vectors A_i on the right or that the vectors A_i annihilate $\{X\}$ on the left. Our question is now: When is a left vector ideal over $\mathfrak{A}$ determined by its right annihilators? In general the totality of left annihilators of $\{X\}$ forms a left vector ideal $\bar{S}_l$ which contains S_l but may also contain other vectors. Hence unless $\bar{S}_l = S_l$ the solutions $\{X\}$ of equations (1.1) do not determine the equations, i.e. S_l, unambiguously since all equations from $\bar{S}_l$ would do equally well.

DEFINITION 1. *The totality of left annihilators S^{rl} of the totality of right annihilators S^r of a set of vectors S is called the left closure $\bar{S}_l$ of S.*

We define the right closure as above, interchanging "right" and "left."

DEFINITION 2. *An algebra $\mathfrak{A}$ is said to be strongly closed if every left (right) vector ideal over $\mathfrak{A}$ is its own left (right) closure.*

DEFINITION 3. *An algebra $\mathfrak{A}$ is said to be weakly closed if every left (right) ideal of $\mathfrak{A}$ is its own left (right) closure.*

It was shown by the author in the above paper that a semi-simple algebra is strongly closed. The result has been announced[2] by J. von Neumann that a regular ring, the infinite analogue of a semi-simple algebra, is strongly closed.

Weak closure of an algebra is an immediate consequence of strong closure, since the ideals of the algebra may be considered as vector ideals of length one.

This paper is devoted to an investigation of the properties of this closure and the characteristics of closed algebras. The two principal results are (1) that a weakly closed algebra has a unit (Theorem 3.2) and (2) that an algebra which is weakly closed is also strongly closed (Theorem 5.2). Hence "strong" closure

[1] Group rings and extensions. I. Annals of Math., vol. 39 (1938), pp. 220–234.

[2] J. von Neumann: "On regular rings." Proc. of the Nat. Acad. of Sci., vol. 22 (1936), pp. 707–713.

is one of the principal properties of "weakly" closed algebras. Various structure (lattice) properties of annihilators and closures are proved in section four of which Theorem 4.5 is the most powerful. Note that in an arbitrary algebra the cross-cut of closed vector ideals is closed, but that the union of closed vector ideals need not be closed. Finally in section six it is shown that an algebra is closed if and only if the ideals in its radical are closed, and this incidentally reproves the closure of semi-simple algebras.

Most of the theorems in this paper are stated for left ideals. It is to be understood that by interchanging "right" and "left" the corresponding theorem for right ideals holds.

2. General Closure Properties

With any set of vectors S over an algebra $\mathfrak{A}$ we have associated a set $\overline{S}_r$ which we call the right closure and a set $\overline{S}_l$ which we call the left closure. To justify this terminology we show that these closures satisfy the general closure postulates[3]

C1 $\overline{S} \supset S$.

C2 $\overline{\overline{S}} = \overline{S}$.

C3 *If* $S \supset T$, *then* $\overline{S} \supset \overline{T}$.

Proof of C1. Let $S = \{X\}$. If $S^r = \{Y\}$, then every vector of $\{X\}$ annihilates every vector of $\{Y\}$ on the left. Hence $\{X\}$ is contained in S^{rl} the totality of left annihilators of $\{Y\}$.

Proof of C2. $S^{rlr} = (S^r)^{lr} = \overline{S_r^r}$ by definition and hence by C1 $S^{rlr} \supset S^r$. On the other hand S^{rlr} annihilates S^{rl} on the right and in particular $S \subset S^{rl}$. Hence $S^{rlr} \subset S^r$ the totality of right annihilators of S. Consequently $S^{rlr} = S^r$ and $S^{rlrl} = S^{rl}$ or $\overline{S}_l = \overline{S}_l$.

Proof of C3. If $S \supset T$, then $S^r \subset T^r$ and from this $S^{rl} \supset T^{rl}$.

In proving C2 we actually proved the stronger statement

Theorem 2.1. $S^{rlr} = S^r$, $S^{lrl} = S^l$.

This states that every left (right) annihilator is its own left (right) closure. As was remarked to the author by J. von Neumann, the closed vector ideals are precisely those which are annihilators. For conversely if $\overline{S}_l = S$, then $S = S^{rl} = (S^r)^l$.

Theorem 2.2. *The right (left) annihilators of a set of vectors over* $\mathfrak{A}$ *form a right (left) vector ideal.*

This is an immediate consequence of the linearity of equations (1.1).

3. Existence of a unit in weakly closed algebras

In this section, save in the statement of the two theorems, where precision dictates otherwise, "closed" will be used instead of "weakly closed," since no reference is made to strong closure.

[3] Garrett Birkhoff: "The meaning of completeness." Annals of Math., Series II, vol. 38, pp. 57-60 (1937).

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

A TYPE OF ALGEBRAIC CLOSURE

BY MARSHALL HALL

(Received July 21, 1938)

1. Introduction

In a previous paper[1] the author was led to the following question: Given the complete set of solutions $\{X\} = [x_1, x_2, \cdots, x_n]$ of linear equations

$$L_i = a_1^{(i)}x_1 + \cdots + a_n^{(i)}x_n = 0, \qquad i = 1, \cdots, s, \tag{1.1}$$

over an algebra $\mathfrak{A}$, to what extent can we determine the equations? Naturally the solutions $\{X\}$ will also satisfy any equation $\sum\lambda_i L_i = 0$ derivable from (1.1) with any λ_i from $\mathfrak{A}$. Write $A_i = [a_1^{(i)}, \cdots, a_n^{(i)}]$. Then $S_l = \sum\lambda_i A_i$, $\lambda_i \subset \mathfrak{A}$ is a left vector ideal over $\mathfrak{A}$ and the solutions $\{X\}$ belong to all linear equations whose coefficients are vectors of S_l. We shall express equations (1.1) by saying that the vectors $\{X\}$ annihilate the vectors A_i on the right or that the vectors A_i annihilate $\{X\}$ on the left. Our question is now: When is a left vector ideal over $\mathfrak{A}$ determined by its right annihilators? In general the totality of left annihilators of $\{X\}$ forms a left vector ideal $\bar{S}_l$ which contains S_l but may also contain other vectors. Hence unless $\bar{S}_l = S_l$ the solutions $\{X\}$ of equations (1.1) do not determine the equations, i.e. S_l, unambiguously since all equations from $\bar{S}_l$ would do equally well.

DEFINITION 1. *The totality of left annihilators S^{rl} of the totality of right annihilators S^r of a set of vectors S is called the left closure $\bar{S}_l$ of S.*

We define the right closure as above, interchanging "right" and "left."

DEFINITION 2. *An algebra $\mathfrak{A}$ is said to be strongly closed if every left (right) vector ideal over $\mathfrak{A}$ is its own left (right) closure.*

DEFINITION 3. *An algebra $\mathfrak{A}$ is said to be weakly closed if every left (right) ideal of $\mathfrak{A}$ is its own left (right) closure.*

It was shown by the author in the above paper that a semi-simple algebra is strongly closed. The result has been announced[2] by J. von Neumann that a regular ring, the infinite analogue of a semi-simple algebra, is strongly closed.

Weak closure of an algebra is an immediate consequence of strong closure, since the ideals of the algebra may be considered as vector ideals of length one.

This paper is devoted to an investigation of the properties of this closure and the characteristics of closed algebras. The two principal results are (1) that a weakly closed algebra has a unit (Theorem 3.2) and (2) that an algebra which is weakly closed is also strongly closed (Theorem 5.2). Hence "strong" closure

[1] Group rings and extensions. I. Annals of Math., vol. 39 (1938), pp. 220–234.

[2] J. von Neumann: "On regular rings." Proc. of the Nat. Acad. of Sci., vol. 22 (1936), pp. 707–713.

is one of the principal properties of "weakly" closed algebras. Various structure (lattice) properties of annihilators and closures are proved in section four of which Theorem 4.5 is the most powerful. Note that in an arbitrary algebra the cross-cut of closed vector ideals is closed, but that the union of closed vector ideals need not be closed. Finally in section six it is shown that an algebra is closed if and only if the ideals in its radical are closed, and this incidentally reproves the closure of semi-simple algebras.

Most of the theorems in this paper are stated for left ideals. It is to be understood that by interchanging "right" and "left" the corresponding theorem for right ideals holds.

2. General Closure Properties

With any set of vectors S over an algebra $\mathfrak{A}$ we have associated a set $\overline{S}_r$ which we call the right closure and a set $\overline{S}_l$ which we call the left closure. To justify this terminology we show that these closures satisfy the general closure postulates[3]

C1 $\overline{S} \supset S$.

C2 $\overline{\overline{S}} = \overline{S}$.

C3 *If* $S \supset T$, *then* $\overline{S} \supset \overline{T}$.

Proof of C1. Let $S = \{X\}$. If $S^r = \{Y\}$, then every vector of $\{X\}$ annihilates every vector of $\{Y\}$ on the left. Hence $\{X\}$ is contained in S^{rl} the totality of left annihilators of $\{Y\}$.

Proof of C2. $S^{rlr} = (S^r)^{lr} = \overline{S_r^r}$ by definition and hence by C1 $S^{rlr} \supset S^r$. On the other hand S^{rlr} annihilates S^{rl} on the right and in particular $S \subset S^{rl}$. Hence $S^{rlr} \subset S^r$ the totality of right annihilators of S. Consequently $S^{rlr} = S^r$ and $S^{rlrl} = S^{rl}$ or $\overline{S}_l = \overline{S}_l$.

Proof of C3. If $S \supset T$, then $S^r \subset T^r$ and from this $S^{rl} \supset T^{rl}$.

In proving C2 we actually proved the stronger statement

Theorem 2.1. $S^{rlr} = S^r$, $S^{lrl} = S^l$.

This states that every left (right) annihilator is its own left (right) closure. As was remarked to the author by J. von Neumann, the closed vector ideals are precisely those which are annihilators. For conversely if $\overline{S}_l = S$, then $S = S^{rl} = (S^r)^l$.

Theorem 2.2. *The right (left) annihilators of a set of vectors over* $\mathfrak{A}$ *form a right (left) vector ideal.*

This is an immediate consequence of the linearity of equations (1.1).

3. Existence of a unit in weakly closed algebras

In this section, save in the statement of the two theorems, where precision dictates otherwise, "closed" will be used instead of "weakly closed," since no reference is made to strong closure.

[3] Garrett Birkhoff: "The meaning of completeness." Annals of Math., Series II, vol. 38, pp. 57-60 (1937).

THEOREM 3.1. *A nilpotent algebra is not weakly closed.*

Let $\mathfrak{A}$ be a nilpotent algebra and suppose $\mathfrak{A}^{n-1} \neq 0$, $\mathfrak{a}^n = 0$. The right annihilator of $(0)_l$ is $\mathfrak{A}$ but the left annihilator of $\mathfrak{A}$ includes $\mathfrak{A}^{n-1}$ and hence is not $(0)_l$. Hence $\mathfrak{A}$ is not closed.

THEOREM 3.2. *A weakly closed algebra $\mathfrak{A}$ has a unit.*

Suppose $\mathfrak{A}$ is a closed algebra. Then by Theorem 3.1 $\mathfrak{A}$ is not nilpotent. An algebra which is not nilpotent must contain an idempotent e. We use the two sided Pierce decomposition of $\mathfrak{A}$ with respect to e

$$x = exe + (xe - exe) + (ex - exe) + (exe - xe - ex + x). \tag{3.1}$$

Here

$$x = x_1 + x_2 + x_3 + x_4, \tag{3.2}$$

where

$$\begin{aligned} &x_1e = x_1, \quad x_2e = x_2, \quad x_3e = 0, \quad x_4e = 0, \\ &ex_1 = x_1, \quad ex_2 = 0, \quad ex_3 = x_3, \quad ex_4 = 0. \end{aligned} \tag{3.3}$$

The corresponding decomposition of $\mathfrak{A}$ is

$$\mathfrak{A} = \mathfrak{A}_1 + \mathfrak{A}_2 + \mathfrak{A}_3 + \mathfrak{A}_4, \tag{3.4}$$

where (Deuring, Algebren, p. 13) any element x of $\mathfrak{A}$ has an unique representation (3.2) with x_i from $\mathfrak{A}_i$, $i = 1, 2, 3, 4$. Now $\mathfrak{A}_4$ is a subalgebra of $\mathfrak{A}$, since $\mathfrak{A}_4\mathfrak{A}_4 \subseteq \mathfrak{A}_4$ being annihilated by e on both sides. Unless $\mathfrak{A}_4$ is nilpotent, it contains an idempotent e_4. If we put $e^* = e + e_4$, e^* is an idempotent and we may obtain a decomposition (3.4) of $\mathfrak{A}$ with respect to e^* in which the new $\mathfrak{A}_1$ contains the old $\mathfrak{A}_1$ and also at least e_4. In this way we may change the idempotent e so as to enlarge $\mathfrak{A}_1$ as much as possible. Let us suppose this done in (3.1)–(3.4). Then $\mathfrak{A}_4$ is nilpotent.

$$\mathfrak{A}_4^n = 0, \qquad n \geqq 1. \tag{3.5}$$

Now if $\mathfrak{A}_2 = \mathfrak{A}_3 = \mathfrak{A}_4 = 0$, $\mathfrak{A} = \mathfrak{A}_1$ and e is a unit of $\mathfrak{A}$ and the theorem is true. We shall prove the theorem by showing that if $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$ do not all vanish then there is an ideal in $\mathfrak{A}$ which is not its own closure and that therefore $\mathfrak{A}$ is not closed.

Suppose that $\mathfrak{A}_2 = \mathfrak{A}_3 = 0$ and $\mathfrak{A}_4 \neq 0$. Then as $\mathfrak{A}_1\mathfrak{A}_4 = 0$ and $\mathfrak{A}_4^n = 0$, $\mathfrak{A}_4^{n-1}$ annihilates $\mathfrak{A}$ on the right and $(0)_r$ is not its own right closure. Hence $\mathfrak{A}$ is not closed. Suppose $\mathfrak{A}_2 = 0$. If $\mathfrak{A}_4 = 0$, then $\mathfrak{A}_3 \neq 0 \subset \overline{(0)}_l$. If $\mathfrak{A}_4 \neq 0$, then $\mathfrak{A}_4^{n-1} \neq 0 \subseteq \overline{(0)}_l$. Hence $(0)_l \neq \overline{(0)}_l$ and $\mathfrak{A}$ is not closed. Similarly if $\mathfrak{A}_3 = 0$, $(0)_r \neq \overline{(0)}_r$ and $\mathfrak{A}$ is not closed.

There remains the possibility that $\mathfrak{A}_2 \neq 0$, $\mathfrak{A}_3 \neq 0$, while $\mathfrak{A}_4$ may or may not vanish. Consider the complex

$$Z = [\mathfrak{A}_2, \mathfrak{A}_3, \mathfrak{A}_4]. \tag{3.6}$$

LEMMA. *If* $\mathfrak{A}_4^n = 0$, $n \geqq 1$, *then* $Z^{2n+1} = 0$.

Z^{2n+1} consists of complexes

$$\Pi = \mathfrak{A}^{(1)}\mathfrak{A}^{(2)} \cdots \mathfrak{A}^{(2n+1)}, \tag{3.7}$$

where $\mathfrak{A}^{(\lambda)}$ is any one of $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$. Because of (3.3) the following relations hold

$$\begin{aligned} \mathfrak{A}_2^2 &= 0, \\ \mathfrak{A}_2\mathfrak{A}_3 &\subset \mathfrak{A}_4, \\ \mathfrak{A}_2\mathfrak{A}_4 &= 0. \end{aligned} \tag{3.8}$$

Hence a product (3.7) not vanishing because of (3.8) contains either (α) $n+1$ $\mathfrak{A}_2$'s and this where $\mathfrak{A}^{(1)} = \mathfrak{A}^{(3)} = \cdots = \mathfrak{A}^{(2n+1)} = \mathfrak{A}_2$ or (β) at most n $\mathfrak{A}_2$'s. From $\mathfrak{A}_2\mathfrak{A}_3 \subset \mathfrak{A}_4$

$$\Pi \subset \mathfrak{A}'^{(1)} \cdots \mathfrak{A}'^{(u)} = \Pi' \tag{3.9}$$

where $u \geqq n+1$ and $\mathfrak{A}'^{(i)} = \mathfrak{A}_3$, $\mathfrak{A}_4$ except possibly that $\mathfrak{A}'^{(u)} = \mathfrak{A}_2$. Now as

$$\begin{aligned} \mathfrak{A}_3^2 &= 0, \\ \mathfrak{A}_4\mathfrak{A}_3 &= 0, \end{aligned} \tag{3.10}$$

a product Π' is zero if an $\mathfrak{A}_3$ is anywhere except at the beginning. Hence the Π's not vanishing because of (3.10) are of one of the forms

$$\mathfrak{A}_4^u, \quad \mathfrak{A}_3\mathfrak{A}_4^{u-1}, \quad \mathfrak{A}_4^{u-1}\mathfrak{A}_2, \quad \mathfrak{A}_3\mathfrak{A}_4^{u-2}\mathfrak{A}_2. \tag{3.11}$$

Since $\mathfrak{A}_4^n = 0$ and $u \geqq n+1$ these all vanish except possibly $\mathfrak{A}_3\mathfrak{A}_4^{u-2}\mathfrak{A}_2$. But here necessarily $\mathfrak{A}^{(1)} = \mathfrak{A}_3$, $\mathfrak{A}^{(2n+1)} = \mathfrak{A}_2$ and in Π there are at most $n-1$ $\mathfrak{A}_2$'s whence $u \geqq n+2$ and so also $\mathfrak{A}_3\mathfrak{A}_4^{u-2}\mathfrak{A}_2 = 0$. This establishes the Lemma.

Let

$$Z^{t-1} \neq 0, \qquad Z^t = 0, \tag{3.12}$$

where by the Lemma t is at most $2n+1$. $(e)_r = \mathfrak{A}_1 + \mathfrak{A}_3$ is annihilated on the left by $\mathfrak{A}_3 + \mathfrak{A}_4$; $(e)_l = \mathfrak{A}_1 + \mathfrak{A}_2$ is annihilated on the right by $\mathfrak{A}_2 + \mathfrak{A}_4$. As Z^{t-1} annihilates $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$ both on the right and on the left, it is contained in the closure of both $(e)_r$ and $(e)_l$. Consequently unless $Z^{t-1} \subset \mathfrak{A}_1$ one of these is not its own closure and $\mathfrak{A}$ is not closed. Let us suppose $Z^{t-1} \subset \mathfrak{A}_1$. Then there must be an $s \leqq t-2$ such that $Z^s \not\subset \mathfrak{A}_1$, $Z^r \subset \mathfrak{A}_1$ for $r > s$.

Note that a product $\mathfrak{A}^{(1)} \cdots \mathfrak{A}^{(s)} \subset Z^s$ will be in one of $\mathfrak{A}_1$, $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$ depending upon $\mathfrak{A}^{(1)}$ and $\mathfrak{A}^{(s)}$ and the relations (3.3). With z_i, $i = 1, 2, 3, 4$, from Z^s the following relations must hold

$$\begin{aligned} z_i\mathfrak{A}_j &= 0, && j = 2, 3, 4 \text{ except possibly } z_3\mathfrak{A}_2 \neq 0, \\ \mathfrak{A}_j z_i &= 0, && j = 2, 3, 4 \text{ except possibly } \mathfrak{A}_3 z_2 \neq 0. \end{aligned} \tag{3.13}$$

Some of these relations are consequences of (3.3). As an example of the others since $z_2\mathfrak{A}_3 \subset Z^{s+1} \subset \mathfrak{A}_1$ necessarily $z_2\mathfrak{A}_3 = z_2\mathfrak{A}_3 e = 0$. Here the z_4 is contained

in the closures of $(e)_r$ and $(e)_l$ and so unless $z_4 = 0$ these ideals are not their own closures. Let us suppose $z_4 = 0$. Hence in Z^s there are elements z_2 or z_3 or both and possibly elements z_1.

In $Z^{s+1} \neq 0$ we have $z_3\mathfrak{A}_2$ and $\mathfrak{A}_3 z_2$ (z_3, z_2 from Z^s) as the only sets containing non vanishing elements. This $z_3 = a_3 \cdots a_4$ or $a_3 \cdots b_3$ and $z_2 = a_2 \cdots b_2$ or $a_4 \cdots a_2$ (a_i, b_i from $\mathfrak{A}_i$). Hence in Z^{s+1} the only non-vanishing elements are of the form $a_3\begin{Bmatrix} b_2 \\ a_4 \end{Bmatrix} \cdots \begin{Bmatrix} b_4 \\ b_3 \end{Bmatrix} a_2$. These may be written alternatively as $a_3 z_2$ or $z_3 a_2$ and so Z^s contains both elements z_2 and z_3. $Z^{s+2} = 0$ since otherwise there would be non-vanishing elements beginning with $\begin{Bmatrix} b_2 \\ a_4 \end{Bmatrix}$ or ending with $\begin{Bmatrix} b_4 \\ b_3 \end{Bmatrix}$ in Z^{s+1}.

Now though Z is not an ideal, Z^{s+1} is a two sided ideal in $\mathfrak{A}_1$. Consider Z^{s+1} as a right ideal in $\mathfrak{A}_1$. Its left annihilators include $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$ and $\mathfrak{A}_3\mathfrak{A}_2$ in $\mathfrak{A}_1$.

(1) Suppose there are no further left annihilators of Z^{s+1}.

The right annihilators of $\mathfrak{A}_2$, $\mathfrak{A}_3$, $\mathfrak{A}_4$ and $\mathfrak{A}_3\mathfrak{A}_2$ include z_3 from Z^s. Hence $(Z^{s+1})_r$ is not its own right closure and $\mathfrak{A}$ is not closed.

(2) Suppose Z^{s+1} has further left annihilators W_1 (necessarily in $\mathfrak{A}_1$) and that Z^{s+1} is its own right closure. Then $W_1 Z^{s+1} = 0$ by definition or

$$W_1 z_3 \mathfrak{A}_2 = 0 \tag{3.14}$$

with z_3 from Z^s where $W_1 z_3 \neq 0$ since z_3 is not in the right closure of Z^{s+1}. But now in consequence of (3.3), (3.13), and (3.14) $W_1 z_3 \cdot \mathfrak{A} = 0$ and $W_1 z_3 \subset \overline{(0)}_l \neq (0)_l$. Hence $\mathfrak{A}$ is not closed.

4. Structure properties of closure

If S and T are left vector ideals over any associative algebra $\mathfrak{A}$, the following are easily proved:

THEOREM 4.1. *If $S > T$, then $S^r \leqq T^r$.*

THEOREM 4.2. $(S \cup T)^r = S^r \cap T^r$.

THEOREM 4.3. $(S \cap T)^r \geqq S^r \cup T^r$.

The status of the inequality in Theorem 4.1 is given by

THEOREM 4.4. *$\mathfrak{A}$ is strongly closed if and only if $S_l > T_l$ implies $S_l^r < T_l^r$, and $S_r > T_r$ implies $S_r^l < T_r^l$.*

If $S_l > T_l$, then $S_l^r \leqq T_l^r$ always. If $\mathfrak{A}$ is strongly closed we may exclude equality since $S_l^r = T_l^r$ implies $S_l^{rl} = T_l^{rl}$ whence $S_l = T_l$. Conversely if equality is always excluded, $\mathfrak{A}$ is strongly closed. For $S_l^{rl} \geqq S_l$ always and if ever $S_l^{rl} > S_l$ then $S_l^{rlr} < S_l^r$ contrary to Theorem 2.1. Hence $S_l^{rl} = S_l$ and $\mathfrak{A}$ is strongly closed.

With further assumptions Theorem 4.3 may be improved:

THEOREM 4.5. *If S, T, and $S^r \cup T^r$ are closed, then $(S \cap T)^r = S^r \cup T^r$.*

PROOF: Under these assumptions and the use of Theorem 4.2

$$S^r \cup T^r = (S^r \cup T^r)^{lr} = (S^{rl} \cap T^{rl})^r = (S \cap T)^r.$$

THEOREM 4.6. $S \cap T \leqq (S \cap T) \leqq \bar{S} \cap \bar{T}$,

$$S \cup T \leqq \bar{S} \cup \bar{T} \leqq \overline{(S \cup T)}.$$

These relations are direct consequences of Theorems 4.1, 2, 3. From the first of these relations we have

THEOREM 4.7. *If S and T are closed then $S \cap T$ is closed.*

The preceding theorems are perhaps less striking for what they include than for what they seem to lack in completeness. Theorem 4.4 makes clear the necessity of permitting equality in Theorem 4.1 for algebras which are not closed. But a counter example will show that in Theorem 4.3 we cannot conclude that $(S \cap T)^r = S^r \cup T^r$ even when both S and T are closed and that the analogue of Theorem 4.7 does not hold for the union of vector ideals.

Let $\mathfrak{A}$ be the algebra over the rational field with a basis 1, e, x where $e^2 = e$, $ex = x$, $xe = 0$, $x^2 = 0$ and let $S = (e)_l$, $T = (e + x)_l$. Here $(S \cap T)^r = (1)_r$ while $S^r \cup T^r = (1 - e, x)_r$ although both S and T are closed. Moreover $S^r = (1 - e)_r$ and $T^r = (1 - e - x)_r$ are both closed but $S^r \cup T^r$ is not closed.

$$\overline{(S^r \cup T^r)} = (1)_r.$$

5. Strong closure a consequence of weak closure

THEOREM 5.1. *If $\mathfrak{A}$ is weakly closed then the principal left vector ideal $S_l = [a_1, a_2, \cdots, a_n]$ is its own left closure.*

The proof will be by induction on n. For $n = 1$ S_l is a principal ideal of $\mathfrak{A}$ and its closure is a special case of the weak closure of $\mathfrak{A}$. Let $n \geqq 2$ and assume that every principal vector ideal of length $n - 1$ is closed.

S_l^r includes those vectors $S_1^r = [x_1', 0, \cdots, 0]$ for every x_1' such that

$$a_1 x_1' = 0. \tag{5.1}$$

S_l^r also includes all those vectors $S^{*r} = [0, x_2^*, \cdots, x_n^*]$ such that

$$a_2 x_2^* + \cdots + a_n x_n^* = 0. \tag{5.2}$$

Let $U = [u_1, u_2, \cdots, u_n]$ be any vector of $\bar{S}_l$. Then for every x_1' of (5.1)

$$u_1 x_1' = 0, \tag{5.3}$$

whence $u_1 \subset \overline{(a_1)}_l$ and from the weak closure of $\mathfrak{A}$, $u_1 = ta_1$. Here $U - t[a_1, \cdots, a_n] = [0, v_2, \cdots, v_n] = V \subset \bar{S}_l$. Now for every $[x_2^*, \cdots, x_n^*]$ of (5.2) we must have

$$v_2 x_2^* + \cdots + v_n x_n^* = 0, \tag{5.4}$$

whence by the closure of principal vector ideals of length $n - 1$ $[v_2, v_3, \cdots, v_n] = [\lambda a_2, \lambda a_3, \cdots, \lambda a_n]$ and $[0, v_2, v_3, \cdots, v_n] = [0, \lambda a_2, \lambda a_3, \cdots, \lambda a_n]$. But the difficult part of the proof is still to come. For every $w \subset (a_1)_r \cap (a_2, \cdots, a_n)_r = \mathfrak{a}_r \cap \mathfrak{b}_r$, $w = a_1 x_1 = -(a_2 x_2 + \cdots + a_n x_n)$ and conversely for every $[x_1, x_2, \cdots, x_n] \subset S_l^r$, $a_1 x_1 = -(a_2 x_2 + \cdots + a_n x_n) \subset \mathfrak{a}_r \cap \mathfrak{b}_r$. Since

$V = [0, \lambda a_2, \cdots, \lambda a_n] \subset \bar{S}_l$, $0 = \lambda a_2 x_2 + \cdots + \lambda a_n x_n = -\lambda a_1 x_1$. Hence $\lambda \subset (\mathfrak{a}_r \cap \mathfrak{b}_r)^l$. From the weak closure of $\mathfrak{A}$, $\mathfrak{a}_r$, $\mathfrak{b}_r$, and $\mathfrak{a}_r^l \cup \mathfrak{b}_r^l$ are all closed and we may obtain $\lambda \subset \mathfrak{a}_r^l \cup \mathfrak{b}_r^l$ by application of Theorem 4.5. But this means that $\lambda = \mu + \nu$ with $\mu \subset \mathfrak{a}_r^l$, $\nu \subset \mathfrak{b}_r^l$ whence $\mu a_1 = 0$, $\nu a_2 = 0, \cdots, \nu a_n = 0$. Hence $V = [0, \lambda a_2, \cdots, \lambda a_n] = [\mu a_1, \mu a_2, \cdots, \mu a_n]$ and $U = t[a_1, \cdots, a_n] + V = (t + \mu)[a_1, \cdots, a_n]$ and $U \subset S_l$. But U was an arbitrary vector of $\bar{S}_l$. Hence $\bar{S}_l = S$ and the theorem is proved.

THEOREM 5.2. *If $\mathfrak{A}$ is weakly closed then every vector ideal over $\mathfrak{A}$ is its own closure and $\mathfrak{A}$ is strongly closed.*

The proof will be by induction on the length n of the vector ideals. For $n = 1$ this reduces to weak closure which is assumed.

Let S_l be a left vector ideal of length n and assume that all vector ideals of length $n - 1$ are closed. As a vector ideal over a finite algebra $\mathfrak{A}$, S_l will have a finite basis

$$A_i = [a_{i1}, a_{i2}, \cdots, a_{in}], \qquad i = 1, \cdots, s. \tag{5.5}$$

Let $S_l^* = \{[0, b_2, \cdots, b_n]\}$ be the left vector ideal of all those vectors of S_l whose first component is zero and let $B_l = \{[b_2, b_3, \cdots, b_n]\}$ be the associated left vector ideal of length $n - 1$.

LEMMA 1. *Let $A = [a_1, a_2, \cdots, a_n]$ be any vector of S_l. Then if $\{[x_2^*, \cdots, x_n^*]\}$ is any vector of B_l^r there exists an x_1 such that*

$$a_1 x_1 + a_2 x_2^* + \cdots + a_n x_n^* = 0. \tag{5.6}$$

If $A_l^r = \{[x_1', x_2', \cdots, x_n']\}$ then the totality of vectors $C_r = \{[x_2', \cdots, x_n']\}$ forms a right vector ideal of length $n - 1$, and if $[c_2, \cdots, c_n] \subset C_r^l$, then $[0, c_2, \cdots, c_n] \subset A_l^{rl} = A_l$ by Theorem 5.1 and hence $[c_2, \cdots, c_n] \subset B_l$. Hence $B_l \supset C_r^l$ and $B_l^r \subset C_r^{lr} = C_r$ since C_r is a vector ideal of length $n - 1$. Consequently $[x_2^*, \cdots, x_*^n] \subset B_l^r \subset C_r$ and there must be an x_1 such that $[x_1, x_2^*, \cdots, x_n^*] \subset A_l^r$.

LEMMA 2. *If $[x_2^*, \cdots, x_n^*] \subset B_l^r$ then there is an x_1 such that*

$$[x_1, x_2^*, \cdots, x_n^*] \subset S_l^r.$$

Now $[x_1, x_2^*, \cdots, x_n^*] \subset S_l^r$ if and only if

$$a_{i1} x_1 + a_{i2} x_2^* + \cdots + a_{in} x_n^* = 0, \qquad i = 1, \cdots, s. \tag{5.7}$$

Lemma 1 assures us that for any individual one of these there will exist an x_1. But we must find an x_1 satisfying all of these simultaneously. Suppose we have found an $x_1^{(j-1)}$ such that

$$a_{i1} x_1^{(j-1)} + a_{i2} x_2^* + \cdots + a_{in} x_n^* = 0, \quad i = 1, \cdots, j - 1, \tag{5.8}$$

and let

$$a_{j1} x_1' + a_{j2} x_2^* \cdots + a_{jn} x_n^* = 0. \tag{5.9}$$

For any $\rho \subset (a_{j,1})_l \cap (a_{1,1}, \cdots, a_{j-1,1})_l$ $\rho = \lambda_1 a_{11} + \cdots + \lambda_{j-1} a_{j-1,1} = \lambda_j a_{j,1}$.

Hence $\lambda_1 A_1 + \cdots + \lambda_{j-1}A_{j-1} - \lambda_j A_j = [0, b_2, \cdots, b_n]$ where $[b_2, \cdots, b_n] \subset B_l$. Consequently

$$(5.10)\quad \lambda_1 a_{12}x_2^* + \cdots + \lambda_1 a_{1n}x_n^* + \cdots + \lambda_{j-1}a_{j-1,2}x_2^* \cdots + \lambda_{j-1}a_{j-1,n}x_n^* = \lambda_j a_{j2}x_2^* \cdots + \lambda_j a_{j,n}x_n^*,$$

and we have from (5.8, 9, 10)

$$(5.11)\quad \rho x_1^{(j-1)} = \rho x_1',$$

or

$$(5.12)\quad \rho(x_1^{(j-1)} - x_1') = 0.$$

Hence $x_1^{(j-1)} - x_1' \subset [(a_{j1})_l \cap (a_{11}, \cdots, a_{j-1,1})_l]^r = (a_{j1})_l^r \cup (a_{11}, \cdots, a_{j-1,1})_l^r$ by Theorem 4.5 because of the weak closure of $\mathfrak{A}$. This means that

$$(5.13)\quad x_1^{(j-1)} - x_1' = \mu + \nu,$$

where $\mu \subset (a_{j,1})_l^r$, $\nu \subset (a_{11}, \cdots, a_{j-1,1})_l^r$ or

$$(5.14)\quad a_{j1}\mu = 0, \qquad a_{i1}\nu = 0 \qquad i = 1, \cdots, j-1.$$

Now put

$$(5.15)\quad x_1^{(j)} = x_1^{(j)} - \nu = x_1' + \mu,$$

and we have in consequence of (5.8), (5.9), and (5.14)

$$(5.16)\quad a_{i1}x_1^{(j)} + a_{i2}x_2^* + \cdots + a_{in}x_n^* = 0, \qquad i = 1, \cdots, j.$$

In this way we may continue until we find an x_1 satisfying all of equations (5.7) and the Lemma is proved.

Using these two lemmas, we may prove the main theorem without much difficulty. Let $U = [u_1, u_2, \cdots, u_n]$ be any vector of $\bar{S}_l = S_l^{rl} \cdot S_l^r$ contains all those vectors $[x_1, 0, \cdots, 0]$ such that

$$(5.17)\quad a_{i1}x_1 = 0, \qquad (i - 1, 2, \cdots, 3).$$

Then

$$(5.18)\quad \begin{aligned} &x_1 \in ((a_{11})_l \cup (a_{21})_l \cup \cdots \cup (a_{s1})_l)^r, \\ &u_1 \in ((a_{11})_l \cup (a_{21})_l \cup \cdots \cup (a_{s1})_l)^{rl} = (a_{11})_l \cup \cdots \cup (a_{s1})_l, \\ &u_1 = \lambda_1 a_{11} + \cdots + \lambda_s a_{s1}. \end{aligned}$$

S_l contains $\lambda_1 S_1 + \cdots + \lambda_s S_s = [u_1, u_2', \cdots, u_n']$. Hence S_l contains a vector $U' = [u_1, u_2', \cdots, u_n']$ with first component u_1. Here $V = U - U' = [0, v_2, \cdots, v_n] \subset \bar{S}_l$ and

$$(5.19)\quad v_2x_2 + \cdots + v_nx_n = 0$$

if $[x_1, x_2, \cdots, x_n]$ is any vector of S_l^r. The totality of $[x_2, x_3, \cdots, x_n] = D_r$ appearing in (5.19) must also satisfy

$V = [0, \lambda a_2, \cdots, \lambda a_n] \subset \bar{S}_l$, $0 = \lambda a_2 x_2 + \cdots + \lambda a_n x_n = -\lambda a_1 x_1$. Hence $\lambda \subset (\mathfrak{a}_r \cap \mathfrak{b}_r)^l$. From the weak closure of $\mathfrak{A}$, $\mathfrak{a}_r$, $\mathfrak{b}_r$, and $\mathfrak{a}_r^l \cup \mathfrak{b}_r^l$ are all closed and we may obtain $\lambda \subset \mathfrak{a}_r^l \cup \mathfrak{b}_r^l$ by application of Theorem 4.5. But this means that $\lambda = \mu + \nu$ with $\mu \subset \mathfrak{a}_r^l$, $\nu \subset \mathfrak{b}_r^l$ whence $\mu a_1 = 0$, $\nu a_2 = 0, \cdots, \nu a_n = 0$. Hence $V = [0, \lambda a_2, \cdots, \lambda a_n] = [\mu a_1, \mu a_2, \cdots, \mu a_n]$ and $U = t[a_1, \cdots, a_n] + V = (t + \mu)[a_1, \cdots, a_n]$ and $U \subset S_l$. But U was an arbitrary vector of $\bar{S}_l$. Hence $\bar{S}_l = S$ and the theorem is proved.

THEOREM 5.2. *If $\mathfrak{A}$ is weakly closed then every vector ideal over $\mathfrak{A}$ is its own closure and $\mathfrak{A}$ is strongly closed.*

The proof will be by induction on the length n of the vector ideals. For $n = 1$ this reduces to weak closure which is assumed.

Let S_l be a left vector ideal of length n and assume that all vector ideals of length $n - 1$ are closed. As a vector ideal over a finite algebra $\mathfrak{A}$, S_l will have a finite basis

$$A_i = [a_{i1}, a_{i2}, \cdots, a_{in}], \qquad i = 1, \cdots, s. \tag{5.5}$$

Let $S_l^* = \{[0, b_2, \cdots, b_n]\}$ be the left vector ideal of all those vectors of S_l whose first component is zero and let $B_l = \{[b_2, b_3, \cdots, b_n]\}$ be the associated left vector ideal of length $n - 1$.

LEMMA 1. *Let $A = [a_1, a_2, \cdots, a_n]$ be any vector of S_l. Then if $\{[x_2^*, \cdots, x_n^*]\}$ is any vector of B_l^r there exists an x_1 such that*

$$a_1 x_1 + a_2 x_2^* + \cdots + a_n x_n^* = 0. \tag{5.6}$$

If $A_l^r = \{[x_1', x_2', \cdots, x_n']\}$ then the totality of vectors $C_r = \{[x_2', \cdots, x_n']\}$ forms a right vector ideal of length $n - 1$, and if $[c_2, \cdots, c_n] \subset C_r^l$, then $[0, c_2, \cdots, c_n] \subset A_l^{rl} = A_l$ by Theorem 5.1 and hence $[c_2, \cdots, c_n] \subset B_l$. Hence $B_l \supset C_r^l$ and $B_l^r \subset C_r^{lr} = C_r$ since C_r is a vector ideal of length $n - 1$. Consequently $[x_2^*, \cdots, x_n^*] \subset B_l^r \subset C_r$ and there must be an x_1 such that $[x_1, x_2^*, \cdots, x_n^*] \subset A_l^r$.

LEMMA 2. *If $[x_2^*, \cdots, x_n^*] \subset B_l^r$ then there is an x_1 such that*

$$[x_1, x_2^*, \cdots, x_n^*] \subset S_l^r.$$

Now $[x_1, x_2^*, \cdots, x_n^*] \subset S_l^r$ if and only if

$$a_{i1} x_1 + a_{i2} x_2^* + \cdots + a_{in} x_n^* = 0, \qquad i = 1, \cdots, s. \tag{5.7}$$

Lemma 1 assures us that for any individual one of these there will exist an x_1. But we must find an x_1 satisfying all of these simultaneously. Suppose we have found an $x_1^{(j-1)}$ such that

$$a_{i1} x_1^{(j-1)} + a_{i2} x_2^* + \cdots + a_{in} x_n^* = 0, \quad i = 1, \cdots, j - 1, \tag{5.8}$$

and let

$$a_{j1} x_1' + a_{j2} x_2^* \cdots + a_{jn} x_n^* = 0. \tag{5.9}$$

For any $\rho \subset (a_{j,1})_l \cap (a_{1,1}, \cdots, a_{j-1,1})_l$ $\rho = \lambda_1 a_{11} + \cdots + \lambda_{j-1} a_{j-1,1} = \lambda_j a_{j,1}$.

Hence $\lambda_1 A_1 + \cdots + \lambda_{j-1}A_{j-1} - \lambda_j A_j = [0, b_2, \cdots, b_n]$ where $[b_2, \cdots, b_n] \subset B_l$. Consequently

$$\lambda_1 a_{12}x_2^* + \cdots + \lambda_1 a_{1n}x_n^* + \cdots + \lambda_{j-1}a_{j-1,2}x_2^* \cdots + \lambda_{j-1}a_{j-1,n}x_n^* = \lambda_j a_{j2}x_2^* \cdots + \lambda_j a_{j,n}x_n^*, \tag{5.10}$$

and we have from (5.8, 9, 10)

$$\rho x_1^{(j-1)} = \rho x_1', \tag{5.11}$$

or

$$\rho(x_1^{(j-1)} - x_1') = 0. \tag{5.12}$$

Hence $x_1^{(j-1)} - x_1' \subset [(a_{j1})_l \cap (a_{11}, \cdots, a_{j-1,1})_l]^r = (a_{j1})_l^r \cup (a_{11}, \cdots, a_{j-1,1})_l^r$ by Theorem 4.5 because of the weak closure of $\mathfrak{A}$. This means that

$$x_1^{(j-1)} - x_1' = \mu + \nu, \tag{5.13}$$

where $\mu \subset (a_{j,1})_l^r$, $\nu \subset (a_{11}, \cdots, a_{j-1,1})_l^r$ or

$$a_{j1}\mu = 0, \qquad a_{i1}\nu = 0 \qquad i = 1, \cdots, j-1. \tag{5.14}$$

Now put

$$x_1^{(j)} = x_1^{(j)} - \nu = x_1' + \mu, \tag{5.15}$$

and we have in consequence of (5.8), (5.9), and (5.14)

$$a_{i1}x_1^{(j)} + a_{i2}x_2^* + \cdots + a_{in}x_n^* = 0, \qquad i = 1, \cdots, j. \tag{5.16}$$

In this way we may continue until we find an x_1 satisfying all of equations (5.7) and the Lemma is proved.

Using these two lemmas, we may prove the main theorem without much difficulty. Let $U = [u_1, u_2, \cdots, u_n]$ be any vector of $\bar{S}_l = S_l^{rl} \cdot S_l^r$ contains all those vectors $[x_1, 0, \cdots, 0]$ such that

$$a_{i1}x_1 = 0, \qquad (i - 1, 2, \cdots, 3). \tag{5.17}$$

Then

$$\begin{aligned} &x_1 \in ((a_{11})_l \cup (a_{21})_l \cup \cdots \cup (a_{s1})_l)^r, \\ &u_1 \in ((a_{11})_l \cup (a_{21})_l \cup \cdots \cup (a_{s1})_l)^{rl} = (a_{11})_l \cup \cdots \cup (a_{s1})_l, \\ &u_1 = \lambda_1 a_{11} + \cdots + \lambda_s a_{s1}. \end{aligned} \tag{5.18}$$

S_l contains $\lambda_1 S_1 + \cdots + \lambda_s S_s = [u_1, u_2', \cdots, u_n']$. Hence S_l contains a vector $U' = [u_1, u_2', \cdots, u_n']$ with first component u_1. Here $V = U - U' = [0, v_2, \cdots, v_n] \subset \bar{S}_l$ and

$$v_2x_2 + \cdots + v_nx_n = 0 \tag{5.19}$$

if $[x_1, x_2, \cdots, x_n]$ is any vector of S_l^r. The totality of $[x_2, x_3, \cdots, x_n] = D_r$ appearing in (5.19) must also satisfy

(5.20) $$b_2x_2 + \cdots + b_nx_n = 0$$

for any vector $[b_2, \cdots, b_n]$ of B_l. Hence $D_r \subset B_l^r$. But by Lemma 2 $B_l^r \subset D_r$ and so $D_r = B_l^r$. Now $[v_2, \cdots, v_n] \subset D_r^l = B_l^{rl} = B_l$ since B_l is a vector ideal of length $n - 1$. Hence $V = [0, v_2, \cdots, v_n] \subset S_l^* \subset S_l$ and $U = U' + V \subset S_l$. But U was an arbitrary vector of $\overline{S}_l$. Hence $\overline{S}_l = S_l$ and our theorem is proved.

6. Relation of idempotents and the radical to closure

In semisimple algebras idempotents play an important rôle, and in particular every ideal $\mathfrak{a}$ can be written as the principal ideal of an idempotent, though for a given $\mathfrak{a}$ the idempotent is not in general unique. Let us investigate the rôle which idempotents play in relation to algebras in general and closure.

Let $\mathfrak{R}$ be the radical of $\mathfrak{A}$. Then $\mathfrak{A}^* = \mathfrak{A}/\mathfrak{R}$ is semisimple.

THEOREM 6.1. *If $\mathfrak{a}$ is any left ideal of $\mathfrak{A}$, $(e^*)_l = \mathfrak{a}^* = (\mathfrak{a}, \mathfrak{R})/\mathfrak{R}$ in $\mathfrak{A}^*$, then in $\mathfrak{A}$ there is an idempotent e such that $e/\mathfrak{R} = e^*$ and $\mathfrak{a} = (e)_l \cup r$ where $r \subset \mathfrak{R}(1 - e)$.*

PROOF: Let e_1 be any element of $\mathfrak{a}$ such that $e_1/\mathfrak{R} = e^*$. Then $e_1^2 - e_1 = t \subset \mathfrak{R}$. Put $e_2 = 3e_1^2 - 2e_1^3$. Then $e_2^2 - e_2 = -3t^2 + 4t^3$. Similarly put $e_3 = 3e_2^2 - 2e_2^3$ etc. t belongs to $\mathfrak{R}$ and must be nilpotent. Here each $e_i^2 - e_i$ $(i = 1, 2, \cdots)$ is divisible by a higher power of t than the last and finally $e_m^2 - e_m = 0$. Put $e_m = e$. Now $e_1 \equiv e_2 \equiv \cdots \equiv e_m \equiv e \pmod{\mathfrak{R}}$ and e is an idempotent. Moreover since $e_1 \subset \mathfrak{a}$ and e is a polynomial in e_1 then $e \subset \mathfrak{a}$. Hence $(e)_l \subset \mathfrak{a}$ and as $(e^*) = \mathfrak{a}/\mathfrak{R}$, $e/\mathfrak{R} = e^*$ it follows that $\mathfrak{a} = (e)_l \cup \rho$ with $\rho \subset \mathfrak{R}$. Here $\rho e \subset (e)_l$. Hence $\mathfrak{a} = (e)_l \cup \rho e \cup \rho = (e)_l \cup \rho e \cup \rho(1 - e) = (e)_l \cup \rho(1 - e)$.

THEOREM 6.2. *If e is an idempotent then $\overline{(e)}_l = (e)_l \cup (\bar{0})_l$. If $\mathfrak{A}$ has a unit the principal ideals of idempotents are closed.*[4]

PROOF: $(e)_l = xe$ for all $x \subset \mathfrak{A}$. Hence $y - ey \subset (e)_l^r$ since $xe(y - ey) = 0$. Again if $z \subset (e)_l^r$ in particular $ez = 0$, whence $z = z - ez$. Hence $(e)_l^r = y - ey$, for all y. Let $w \subset \overline{(e)}_l$. $w = we + (w - we)$ where $we \subset (e)_l \subset \overline{(e)}_l$. But $(w - we)(y - ey) = 0$ implies $(w - we)y = 0$, whence $(w - we) \subset (\bar{0})_l$. Hence $\overline{(e)}_l \subset (e)_l \cup (\bar{0})_l$. But $(e)_l \subset \overline{(e)}_l$ and $(\bar{0})_l$ is contained in the closure of any ideal. Hence $\overline{(e)}_l = (e)_l \cup (\bar{0})_l$.

THEOREM 6.3. *Let $\mathfrak{A}$ be an algebra with a unit and e an idempotent. If $\mathfrak{b} = (e)_l \cup \mathfrak{a}_l = (e)_l \cup \mathfrak{a}(1 - e)$, then $\bar{\mathfrak{b}} = (e)_l \cup \overline{(\mathfrak{a}_l(1 - e))}$.*

PROOF: If $x \subset \mathfrak{a}_l$, $x = xe + x(1 - e)$ and $xe \subset (e)_l$. Hence $(e)_l \cup \mathfrak{a}_l = (e)_l \cup \mathfrak{a}_l(1 - e)$. By Theorem 4.2 $\mathfrak{b}^r = (e)_l^r \cup (\mathfrak{a}(1 - e))^r = (1 - e)_r \cap (\mathfrak{a}(1 - e))^r$. But this implies $\mathfrak{b}^r = (1 - e)(\mathfrak{a}(1 - e))^r$. For if $y \subset (1 - e)_r$ then $(1 - e)y = y$, and if $y \subset (\mathfrak{a}(1 - e))^r$ then $y = (1 - e)y \subset (1 - e)(\mathfrak{a}(1 - e))^r$; hence $\mathfrak{b}^r \subset (1 - e)(\mathfrak{a}(1 - e))^r$. Conversely if $z \subset (1 - e)(\mathfrak{a}(1 - e))^r$ then $z \subset (1 - e)_r$ and $(1 - e)z = z$.

Also $z = (1 - e)w$ where w is a right annihilator of $\mathfrak{a}(1 - e)w$. Hence

[4] This theorem comes from a remark of J. von Neumann.

$\mathfrak{a}(1-e)w = 0$, or $(1-e)(1-e)w = 0$, or $(1-e)z = 0$ and $z \subset (\mathfrak{a}(1-e))^r$. Consequently $(1-e)(\mathfrak{a}(1-e))^r \supset \mathfrak{b}^r$ and $\mathfrak{b}^r = (1-e)(\mathfrak{a}(1-e))^r$. Hence $\bar{\mathfrak{b}} = [(1-e)(\mathfrak{a}(1-e))^r]^l$. Now $\bar{\mathfrak{b}} \supset (e)_l$ and if $x \subset \bar{\mathfrak{b}}$, $x = xe + x(1-e)$, whence $z = x(1-e) \subset \bar{\mathfrak{b}}$. If $z(1-e)(\mathfrak{a}(1-e))^r = 0$, then $z(\mathfrak{a}(1-e))^r = 0$, whence $z \subset (\mathfrak{a}(1-e))^{rl} = (\mathfrak{a}(1-e))$. Consequently $\bar{\mathfrak{b}} \subset (e)_l \cup (\mathfrak{a}(1-e))$. But by the second part of Theorem 4.6 $(e)_l \cup (\mathfrak{a}(1-e)) \subset \bar{\mathfrak{b}}$ and by Theorem 6.2 $(\bar{e})_l = (e)_l$. Hence $\bar{\mathfrak{b}} = (e)_l \cup (\mathfrak{a}_l(1-e))$ and the theorem is proved.

THEOREM 6.4. *An algebra $\mathfrak{A}$ with a unit is closed if and only if the ideals in the radical $\mathfrak{R}$ are closed.*

If $\mathfrak{A}$ is closed then it is necessary a fortiori that the ideals in $\mathfrak{R}$ be closed. For sufficiency of this condition suppose that the ideals in $\mathfrak{R}$ are closed. Let $\mathfrak{a}$ be any left ideal of $\mathfrak{A}$. Then by Theorem 6.1 $\mathfrak{a} = (e)_l \cup r$ where $r \subset \mathfrak{R}(1-e)$. By Theorem 6.3 $\bar{\mathfrak{a}} = (e)_l \cup \bar{r}$ whence $\mathfrak{a}$ is closed if r is closed. Hence if the ideals in $\mathfrak{R}$ are closed, then $\mathfrak{A}$ is closed.

YALE UNIVERSITY.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

FUCHSIAN GROUPS AND MIXTURES

BY GUSTAV A. HEDLUND

1. Introduction

A flow in a space is metrically transitive if whenever the space is divided into two complementary, measurable, invariant sets, one of the sets is of measure zero. Metric transitivity implies that for almost all motions the average time spent in measurable sets is the same for sets of equal measure. It is known (Hopf, Ch. V)[1] that the flow defined by the geodesics on any surface of constant negative curvature, of finite connectivity, and of finite (hyperbolic) area, is metrically transitive.

A deeper property than metric transitivity is that which has been called mixture (cf. Hopf, p. 36). If a flow is a mixture, any set of positive measure spreads itself homogeneously throughout the space with increasing or decreasing time. If a flow is a mixture it is metrically transitive.

The object of the present paper is to show that the geodesic flow on a surface of constant negative curvature, of finite connectivity, and of finite area, is a mixture. The derivation is elementary and geometrical. It makes use of the known metrical transitivity of the flow, so it does not furnish a new proof of that property. Essential use is made of the horocycles, which have been studied in a previous paper by the author (cf. Hedlund).

2. The hyperbolic plane and the space E of elements

The interior Ψ of the unit circle U, $z\bar{z} = 1$, provided with the metric

$$ds = \frac{2\,|\,dz\,|}{1 - z\bar{z}}, \qquad z\bar{z} < 1, \tag{2.1}$$

is a hyperbolic plane. The *hyperbolic length* or *H-length* of a curve γ of Class C' in ψ is $\int_\gamma ds$, ds given by (2.1). The geodesics or *H-lines* (*hyperbolic lines*) are arcs of circles orthogonal to U. The *H-distance* (*hyperbolic distance*), $H(P_1, P_2)$, between two points P_1 and P_2 of Ψ is the H-length of the unique H-line segment joining these points.

An *H-circle* (*hyperbolic circle*) is the locus of points at equal H-distances from a fixed point of Ψ and the fixed point will be called the *center* of the H-circle. An H-circle is also a euclidean circle, the euclidean center not necessarily coinciding with the center just defined. Oriented H-circles will be termed *right* or *left* according as the sense of rotation is clockwise or counterclockwise.

[1] Confer bibliography at end of paper for references.

A *horocycle* is a euclidean circle internally tangent to U. The point Q of contact with U is the *point at infinity* of the horocycle and this point together with the euclidean radius completely determine the horocycle. The horocycle with point at infinity Q and euclidean radius r will be denoted by $C(Q, r)$. The horocycle $C(Q, r)$ is an orthogonal trajectory of the set of H-lines having Q as end point. Oriented horocycles will be classified as *right*, $C_R(Q, r)$, or *left*, $C_L(Q, r)$, according as the sense of rotation on the horocycle is clockwise or counterclockwise.

Hyperbolic area is defined by the integral

$$\iint \frac{4dxdy}{(1 - z\bar{z})^2}. \tag{2.2}$$

The metric (2.1) and the integral (2.2) are invariant under linear fractional transformations which take Ψ into itself, so that under such transformations, hyperbolic distance and hyperbolic area, as well as angle, are invariant. These transformations carry H-lines into H-lines, H-circles into H-circles, and horocycles into horocycles.

Let $\bar{E}$ be the set (x, y, ϕ), $x^2 + y^2 < 1$ and $-\infty < \phi < +\infty$, with neighborhoods defined by considering this set as a subset of euclidean space. Let E be the space obtained by considering the points $(x, y, \phi + 2n\pi)$, $n = 0, \pm 1, \cdots$, as identical. If the angle ϕ at each point $P(x, y)$ of Ψ is measured in the clockwise sense from the direction through the point parallel to the positive x-axis, to each point of E there corresponds a point in Ψ and a direction at this point, and conversely. A point in Ψ together with a direction at this point is an *element*, so that E is the space of elements in the hyperbolic plane. The point of Ψ associated with the element is the point *bearing the element*.

Given a point of E, there are an infinite number of points of $\bar{E}$, the identification of which defined the given point. Of this infinite set there is just one such that $0 \leqq \phi < 2\pi$, so that a set of points of E is in one-to-one correspondence with a subset of $\bar{E}$ for which $0 \leqq \phi < 2\pi$. Let the subset of $\bar{E}$ for which $0 \leqq \phi < 2\pi$ be denoted by $\bar{E}^*$. The set E_1 of points in E will be said to be *measurable* if the set $\bar{E}_1^*$ in one-to-one correspondence with it in $\bar{E}^*$ is measurable, considering $\bar{E}^*$ as a subset of the euclidean space (x, y, ϕ). This is equivalent to the condition that the points of $\bar{E}$ corresponding to E_1 be a measurable set. The *measure* of E will be defined as

$$mE_1 = \iiint_{\bar{E}^*} 4(1 - x^2 - y^2)^{-2}\, dx\, dy\, d\phi. \tag{2.3}$$

This method of defining measure is equivalent to that of Hopf (cf. Hopf, p. 68).

Given a set E_1 in E, the set E_2 will be said to be obtained from E_1 by a *rotation* through the angle α if the elements in E_2 are those obtained by rotating each element in E_1 through the angle α, the sense of rotation being clockwise if α is positive, counterclockwise if α is negative. If E_1 is measurable, the same is

evidently true of E_2, and from the definition of measure as given by (2.3), we have $mE_1 = mE_2$.

3. The geodesic, horocycle and circle flows in E

A point e of E, or element, determines a unique directed H-line g having this element as one of its elements. Let the H-length s along g be measured from the point P_0 bearing the given element and be positive in the positive direction on g. Let ${}_ge_s$ denote the element of g at the point determined by s. The transformation G_s^* defined by $e \to {}_ge_s$ is a continuous, one-to-one transformation of E into itself and depends continuously on the parameter s. This is the flow defined by the geodesics or hyperbolic lines and will be called the *geodesic flow*. It is not difficult to show that measure, as defined by (2.3), is invariant under G_s^*.

A point e of E determines a unique right horocycle having e as one of its elements. In a manner analogous to that in which the geodesic flow is defined, a *right horocycle flow* ${}_RH_s^*$ can be defined by using right horocycles instead of hyperbolic lines. The *left horocycle flow* ${}_LH_s^*$ is defined similarly by making use of left horocycles.

It will now be shown that the right and left horocycle flows preserve measure. It is convenient to derive first some elementary geometric results.

Let $\overset{\frown}{PQ}$ be an arc of a horocycle with both P and Q lying in Ψ. Let ϕ_P denote the smallest positive angle at P between the arc $\overset{\frown}{PQ}$ and the H-line segment PQ. Let ϕ_Q be the similarly defined angle at Q. Since horocycles are circles and H-lines are either circular arcs or straight line segments, it follows that $\phi_P = \phi_Q$ and this common value will be denoted by ϕ_{PQ}. Evidently, $0 < \phi_{PQ} < \pi/2$.

LEMMA 3.1. *Given $s > 0$, let $\overset{\frown}{PQ}$ be an arc of H-length s of any horocycle. Then $H(P, Q)$ and ϕ_{PQ} depend only on s and not on the horocycle or the particular arc of it. Denoting $H(P, Q)$ and ϕ_{PQ} by $H(s)$ and $\phi(s)$, respectively, the functions $H(s)$ and $\phi(s)$ are continuous and increasing, and*

$$\lim_{s \to +\infty} H(s) = +\infty, \qquad \lim_{s \to +\infty} \phi(s) = \pi/2.$$

In the proof of this lemma it can be assumed that the horocycle becomes a right horocycle when so oriented that the arc $\overset{\frown}{PQ}$ with P as initial point is positively directed, for if this is not the case, the proof which follows will apply if the rôles of P and Q are interchanged.

The point P can always be transformed to the origin by a linear fractional transformation taking Ψ into Ψ and such that the element at P of the directed arc $\overset{\frown}{PQ}$ is transformed into a given element at the origin. Let $C_R(A, r)$ be the unique right horocycle determined by this element. Since the linear fractional transformation transforms right horocycles into right horocycles, the arc $\overset{\frown}{PQ}$

is always transformed into an arc $\overset{\frown}{OQ'}$ of $C_R(A, r)$ and the point Q' depends only on s. Since $H(P, Q) = H(O, Q')$, $H(P, Q)$ depends only on s. Since the transformation is conformal, ϕ_{PQ} depends only on s. As s becomes infinite the point Q' approaches A (in the euclidean sense) and $\lim_{s\to+\infty} H(s) = +\infty$. As s becomes infinite, the angle $\phi(s)$ approaches the angle between $C_R(A, r)$ and the hyperbolic ray OA. But $C_R(A, r)$ is orthogonal to OQ, so $\lim_{s\to+\infty} \phi(s) = \pi/2$. That $H(s)$ and $\phi(s)$ are continuous and increasing is obvious. The proof of the lemma is complete.

THEOREM 3.1. *The right and left horocycle flows are measure preserving.*

The discussion will be restricted to the case of the right horocycle flow ${}_RH_s^*$, since the proof for the left horocycle flow is entirely similar. Also, only the case $s > 0$ will be considered, as there is no difficulty in applying the proof to the case $s < 0$.

Lemma 3.1 yields a simple decomposition of ${}_RH_s^*$ into rotations and the transformation $G_{H(s)}^*$, all of which are known to preserve measure. For, given e, a point of E, and $s > 0$, let us determine ${}_RH_s(e)$, which we will denote by ${}_Re_s$. Let P be the point bearing e and let C_R be the right horocycle determined by e. Let Q be the point of C_R such that the directed arc $\overset{\frown}{PQ}$, with initial point P, of C_R is positively directed and has H-length s. Then ${}_Re_s$ is the element of C_R at Q. Using the notation of Lemma 3.1, we let $H(P, Q) = H(s)$. It follows from that lemma that the element ${}_Re_s$ can be obtained from e by rotating e through the angle $\phi(s)$, thus determining an element e^*, then passing to the element $G_{H(s)}(e^*) = {}_Oe_{H(s)}^*$, and finally, rotating ${}_Oe_{H(s)}^*$ through the angle $\phi(s)$. (It should be noted that in the case under consideration both the rotations are positive through the angle $\phi(s)$. The only other possibility occurring in the omitted cases is that both rotations are negative through the angle $-\phi(s)$.)

Let R_α^* denote the transformation of E into itself which keeps the point of Ψ associated with each element fixed and rotates the direction of each element through the angle α. The preceding results can be stated symbolically as follows:

$$(3.1) \qquad {}_RH_s^* = R_{\phi(s)}^* \cdot G_{H(s)}^* \cdot R_{\phi(s)}^*, \qquad s > 0,$$

while analogous formulas hold in the other cases. The stated theorem is an immediate consequence of the fact that all the transformations on the right of (3.1) are measure preserving.

Given $\rho > 0$, a point e of E determines a unique right (left) H-circle of radius ρ and having e as one of its elements. Thus a right (left) circle flow ${}_R^*C_s^\rho$ (${}_L^*C_s^\rho$) can be defined in a manner analogous to that in which the geodesic flow is defined by making use of right (left) H-circles of radius ρ instead of geodesics. Since the proof of the following theorem is entirely similar to that of Theorem 3.1, it will be omitted.

THEOREM 3.2. *The right and left circle flows are measure preserving.*

4. The surface M, the space of elements E_M and flows in E_M

Let F be a Fuchsian group with U as principle circle. Such a group has a normal fundamental region R containing the origin and bounded by H-lines or segments of such lines which are congruent in pairs. If points which are congruent under F are considered identical, there is defined a two-dimensional manifold M of constant negative curvature.

The group F defines a group Γ of transformations of E into itself (Hopf, pp. 70–71). Furthermore, the transformations of Γ preserve measure in E. If elements which are congruent under Γ are considered identical, there is defined a space E_M, the space of elements on M. To each point of E_M there corresponds an infinite set of congruent (under Γ) points of E. Of this infinite set there will be at least one with the point bearing it in R. There will be only one with the point bearing it interior to R. Given a set E_M^* in E_M, let E^* be a set in E in one-to-one correspondence with E_M^* such that all points bearing the elements of E^* are in R and such that the element e^* of E^* corresponding to a point of E_M^* is one of the infinite set of E corresponding to this point. It is evident that if one of the sets E^* corresponding to a given E_M^* is measurable, all will be, for any two differ only by a set of measure zero. The set E_M^* of E_M will be *measurable* if the set E^* is measurable and the *measure* of E_M^* will be defined as that of E^*.

The geodesic, horocycle, and circle flows in E transform congruent elements into congruent elements, so that each of these flows defines a flow in E_M which will be called respectively the *geodesic, horocycle* and *circle flows in* E_M and denoted by G_s, ${}_RH_s$ or ${}_LH_s$, ${}_RC_s^\rho$ or ${}_LC_s^\rho$, as the case may be. It is readily shown that these flows in E_M are measure preserving due to the measure preserving properties of G_s^*, ${}_RH_s^*$, ${}_LH_s^*$, ${}_R^*C_s^\rho$ and ${}_L^*C_s^\rho$.

For later purposes it is desirable to introduce a metric into the space E_M. If (x_1, y_1, ϕ_1) and (x_2, y_2, ϕ_2) are points of $\bar{E}$ representing elements e_1 and e_2 of E, the *distance* between these elements will be defined as

$$H(P_1, P_2) + \| \phi_1 - \phi_2 \|,$$

where P_1 is the point (x_1, y_1) of Ψ, P_2 is the point (x_2, y_2) and $\| \phi_1 - \phi_2 \|$ is the minimum of the set $| \phi_1 - \phi_2 + 2n\pi |$, $n = 0, \pm 1, \cdots$. This definition is evidently independent of the choice of points of $\bar{E}$ representing the given elements. If p_1 and p_2 are points of E_M or elements on M, there correspond to p_1 and p_2 infinite sets S_1 and S_2, respectively, in E. Then $D(p_1, p_2)$, the *distance* between p_1 and p_2, will be defined as the greatest lower bound of the distances between pairs of points of which one is in S_1 and the other is in S_2 (cf. Morse, p. 53).

5. The metric transitivity of horocycle flows

The group F will now be restricted to being of the first kind with a finite set of generators. Under these conditions it is known that $mE_M < \infty$ and that the geodesic flow G_s is metrically transitive (cf. Hopf, Ch. V). It will now be

shown that this implies the metric transitivity of the horocycle flow ${}_RH_s$. The proof for ${}_LH_s$ is similar.

To prove the metric transitivity of ${}_RH_s$, it must be shown that any measurable set V of E_M which is invariant under ${}_RH_s$, and thus consists of the elements on a set of right horocycles which is invariant under F, is either a set of measure zero or the complement of such a set. Thus it can be assumed that V is an invariant (under ${}_RH_s$) set of positive measure and it suffices to show that V has points in common with any measurable set W in E_M and of positive measure.

Recalling the notation of §3, R_α^* denotes a transformation of E into itself which takes the element determined by (x, y, ϕ) into that determined by $(x, y, \phi + \alpha)$, and, as was pointed out there, R_α^* is measure preserving. Under this transformation congruent elements transform into congruent elements, so that R_α^* defines a transformation R_α of E_M into itself which is also measure preserving. Let V be a measurable set in E_M and let $R_\alpha(V) = V_\alpha$.

LEMMA 5.1. $\lim_{\alpha \to 0} m(V_\alpha \cdot V) = \lim_{\alpha \to 0} (V_\alpha + V) = mV.$

For let P be a point interior to the fundamental region R and let S_r, $r > 0$, be the interior of a square with center P and with sides of length r such that all points in S_r are in the interior of R. Let E_0 be the set of elements of E_M determined by the set (x, y, ϕ) where the point (x, y) lies in S_r and $0 < \phi_1 < \phi < \phi_2 < 2\pi$. The set E_0 is an open set in E_M.

The lemma is obviously true for an open set of the type E_0. It follows that it is true for an open set consisting of the sum of a finite number of non-overlapping sets of type E_0. The passage to the case of a denumerable set of non-overlapping sets of this type offers no difficulty. An open set in E_M is, except for a set of measure zero, the sum of a denumerable set of non-overlapping sets of the type of E_0, so the lemma holds for open sets. By complementation it holds for closed sets. Since a measurable set contains closed sets of measure arbitrarily close to that of the given set, and is contained in open sets of measure arbitrarily close to that of the given set, the lemma follows for measurable sets.

THEOREM 5.1. *The right and left horocycle flows are metrically transitive.*

Let V and W be measurable sets of positive measure in E_M, with V invariant under ${}_RH_s$. Then $V_{\pi/2}$ and $W_{-\pi/2}$ are measurable sets and $mV_{\pi/2} = mV$, $mW_{-\pi/2} = mW$. The metric transitivity of the geodesic flow G_s implies (cf. von Neumann, pp. 590–591) the following:

$$\lim_{s-t \to +\infty} \frac{1}{s-t} \int_t^s m[G_\tau(V_{\pi/2}) \cdot W_{-\pi/2}]\, d\tau = \frac{mV_{\pi/2} mW_{-\pi/2}}{mE_M}.$$

It follows that there exists an infinite sequence $0 < s_1 < s_2 < \cdots$, with $\lim_{i \to \infty} s_i = +\infty$, such that

$$m[G_{s_i}(V_{\pi/2}) \cdot W_{-\pi/2}] > \frac{mV_{\pi/2} mW_{-\pi/2}}{2mE_M} = \frac{mVmW}{2mE_M}, \quad i = 1, 2, \cdots. \tag{5.1}$$

Let $\bar{s}_i$ be the number such that $s_i = H(\bar{s}_i)$, where $H(s)$ is as defined in §3.

The existence and uniqueness of $\bar{s}_i$ follows from the facts that $H(0) = 0$, $H(s)$ is continuous, and $H(s)$ is increasing with $\lim_{s \to +\infty} H(s) = +\infty$. From these properties of $H(s)$ we also have $\lim_{i \to \infty} \bar{s}_i = +\infty$. Let $\phi(\bar{s}_i) = \phi_i$, $i = 1, 2, \cdots$, where $\phi(s)$ is as defined in §3. Then, with the aid of Lemma 3.1, $\lim_{i \to \infty} \phi_i = \pi/2$.

From

$$V_{\pi/2} = V_{\phi_i} \cdot V_{\pi/2} + (V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2})$$

it follows that

$$G_{s_i}(V_{\pi/2}) = G_{s_i}(V_{\phi_i} \cdot V_{\pi/2}) + G_{s_i}(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2})$$

and hence

$$G_{s_i}(V_{\pi/2}) \cdot W_{-\pi/2} = G_{s_i}(V_{\phi_i} \cdot V_{\pi/2}) \cdot W_{-\pi/2} + G_{s_i}(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}) \cdot W_{-\pi/2}, \tag{5.2}$$

where the two sets on the right have no common point. Since $\lim_{i \to \infty} \phi_i = \pi/2$, with the aid of Lemma 5.1 there exists an N_1 such that

$$m(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}) < \frac{mVmW}{6mE_M}, \qquad i > N_1. \tag{5.3}$$

Since G_s is a measure preserving transformation,

$$mG_{s_i}(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}) = m(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}),$$

and hence, from (5.3),

$$\begin{cases} m[G_{s_i}(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}) \cdot W_{-\pi/2}] \\ \leqq mG_{s_i}(V_{\pi/2} - V_{\phi_i} \cdot V_{\pi/2}) < \dfrac{mVmW}{6mE_M}, \end{cases} \qquad i > N_1. \tag{5.4}$$

Applying (5.1) and (5.4) to (5.2), there results

$$m[G_{s_i}(V_{\phi_i} \cdot V_{\pi/2}) \cdot W_{-\pi/2}] > \frac{mVmW}{3mE_M}, \qquad i > N_1. \tag{5.5}$$

This evidently implies

$$m[G_{s_i}(V_{\phi_i}) \cdot W_{-\pi/2}] > \frac{mVmW}{3mE_M}, \qquad i > N_1. \tag{5.6}$$

Again making use of Lemma 5.1, there exists an integer $N_2 > N_1$ such that

$$m(W_{-\pi/2} - W_{-\phi_i} \cdot W_{-\pi/2}) < \frac{mVmW}{12mE_M}, \qquad i > N_2. \tag{5.7}$$

From

$$W_{-\pi/2} = W_{-\phi_i} \cdot W_{-\pi/2} + (W_{-\pi/2} - W_{-\phi_i} \cdot W_{-\pi/2})$$

and the fact that the sets on the right have no points in common,

(5.8) $m(V_i \cdot W_{-\pi/2}) = m(V_i \cdot W_{-\phi_i} \cdot W_{-\pi/2}) + m\{V_i \cdot (W_{-\pi/2} - W_{-\phi_i} \cdot W_{-\pi/2})\}$,

where $V_i = G_{s_i}(V_{\phi_i})$. By combining (5.6), (5.7) and (5.8) we obtain

$$m(V_i \cdot V_{-\phi_i}) \geqq m(V_i \cdot W_{-\phi_i} \cdot W_{-\pi/2}) > \frac{mVmW}{4mE_M}, \qquad i > N_2. \tag{5.9}$$

From (3.1),

$${}_RH_s = R_{\phi(s)} G_{H(s)} R_{\phi(s)},$$

and thus

$$\begin{cases} {}_RH_{s_i}(V) \cdot W = [R_{\phi_i}(G_{s_i}\{V_{\phi_i}\})] \cdot W \\ \quad = R_{\phi_i}[G_{s_i}(V_{\phi_i}) \cdot W_{-\phi_i}] = R_{\phi_i}[V_i \cdot W_{-\phi_i}]. \end{cases} \tag{5.10}$$

Since V is assumed to be invariant under ${}_RH_s$,

$${}_RH_{s_i}(V) \cdot W = V \cdot W$$

and since $R_{\phi i}$ preserves measure, we have, with the aid of (5.10),

$$m(V \cdot W) = m(V_i \cdot W_{-\phi_i}).$$

Applying (5.9), there results

$$m(V \cdot W) > \frac{mVmW}{4mE_M}, \qquad i > N_2,$$

so that the sets V and W must have points in common. The proof that the right horocycle flow is metrically transitive is complete.

The similar proof that the left horocycle flow is metrically transitive is omitted.

6. The geodesic flow is a mixture

The flow G_s is a mixture if, V and W being arbitrary measurable sets of E,

$$\lim_{s \to \pm\infty} m[G_s(V) \cdot W] = \frac{mVmW}{mE_M}. \tag{6.1}$$

This relationship will hold for measurable sets if it holds for all open sets, and it will hold for all open sets if it holds for a particular set S^* of open sets such that any open set is the sum, except for a set of measure zero, of a denumerable set of non-overlapping sets from S^*. In the following V will be restricted to a set S^* satisfying the stated conditions, while W will be any open set of E_M. The set S^* can be chosen, in particular, as follows. Let σ be the interior of a (euclidean) square lying, together with its boundary, in the interior of the fundamental region R. Let I be an open interval of U. Let $E(\sigma, I)$ be the set of initial elements of directed H-rays (hyperbolic rays) with initial point in σ and terminal point in I. No two elements in $E(\sigma, I)$ are congruent, for no two interior points of R are congruent and there are no fixed points of elliptic transformations interior to R. Thus the elements $E(\sigma, I)$ determine a set O^* of E_M such that the points of O^* are in one-to-one correspondence with the set

$E(\sigma, I)$. The sets O^* evidently form a set S^* of open sets of the desired kind. Thus it is sufficient to prove that if O^* belongs to S^* and O is any open set, then

$$\lim_{s\to\pm\infty} m[G_s(O^*)\cdot O] = \frac{mO^*mO}{mE_M}. \tag{6.2}$$

The proof will be carried through only for the case $s \to +\infty$, inasmuch as the proof is similar for $s \to -\infty$.

Let Q be a point of the square σ used in determining O^* and let $E(Q, I)$ be the set of initial elements of the H-rays with initial point at Q and terminal point in I. Let O_Q^* be the subset of O^* determined by the set $E(Q, I)$. Then the set $G_s(O_Q^*)$ is the set in E_M determined by the elements situated on a segment of an H-circle with center Q and radius s, and with directions perpendicular to the H-circle and directed outward. The segment of the H-circle is that between the geodesic rays with initial points at Q and points at infinity at the ends of the segment I used in determining O^*. Thus the set $[G_s(O^*)]_{\pi/2}$ is the set determined by the elements on a set of arcs of H-circles of radius s and with centers in σ. As s becomes infinite the minimum length of these arcs becomes infinite. Suppose it can be shown that

$$\lim_{s\to+\infty} m\{[G_s(O^*)]_{\pi/2}\cdot O_{\pi/2}\} = \frac{mO^*_{\pi/2}mO_{\pi/2}}{mE_M}, \tag{6.3}$$

where the subscript $\pi/2$ again indicates rotation through the angle $\pi/2$. Then the corresponding relationship in (6.2) holds since measure is preserved by the rotation through the angle $-\pi/2$. For brevity let $[G_s(O^*)]_{\pi/2} = O_s$, $O_{\pi/2} = \bar{O}$, and $E_M = \Omega$. Then the desired relationship is

$$\lim_{s\to+\infty} m(O_s\cdot\bar{O}) = \frac{mO_0m\bar{O}}{m\Omega}. \tag{6.4}$$

Since G_s and rotations preserve measure, $mO_s = mO_0$.

Now consider the set O_s, a set determined by the elements on a set of arcs of right H-circles, each of radius s. The desired result (6.4) is implied by

$$\liminf_{s\to+\infty} m(O_s\cdot\bar{O}) \geqq \frac{mO_0m\bar{O}}{m\Omega}. \tag{6.5}$$

For if (6.5) holds for sets of the stated kind, it holds for measurable sets, and, in particular, for closed sets. Thus

$$\liminf_{s\to+\infty} m(O_s\cdot C_\Omega(\bar{O})) \geqq \frac{mO_0mC_\Omega(\bar{O})}{m\Omega},$$

where $C_\Omega(\bar{O})$ is the complement of $\bar{O}$ with respect to Ω. From

$$m(O_s\cdot\bar{O}) + m(O_s\cdot C_\Omega(\bar{O})) = mO_s = mO_0$$

it follows that

$$\limsup_{s\to+\infty} m(O_s\cdot\bar{O}) = mO_0 - \liminf_{s\to+\infty} m(O_s\cdot C_\Omega(\bar{O})).$$

Thus

$$\limsup_{s\to+\infty} m(O_s\cdot\bar{O}) \leqq mO_0 - \frac{mO_0 mC_\Omega(\bar{O})}{m\Omega} = \frac{mO_0 m\bar{O}}{m\Omega},$$

and (6.4) holds.

Let O_s^1 be the set of $E_M = \Omega$ determined by the elements on the set of all subsegments of the arcs of right H-circles determining O_s such that the initial point of each subsegment is at the initial point of the arc containing it and each subsegment is of length L. Let O_s^2 be the set obtained by repeating the process, but using the terminal points of segments of O_s^1 as initial points. This process can be repeated a certain number of times, say $k(s)$, without leaving the arcs determining O_s. Since the minimum length of the segments of O_s becomes infinite with s, $k(s)$ does the same. Let $O_s^* = \sum_{i=1}^{k(s)} O_s^i$. We omit the elementary proof of the following statement.

$$\lim_{s\to+\infty} mO_s^* = mO_s = mO_0. \tag{6.6}$$

To prove (6.5) it is sufficient to show that, given $\epsilon > 0$, there exists an S such that

$$m(O_s^*\cdot\bar{O}) \geqq \frac{m\bar{O}}{m\Omega} mO_s^* - \epsilon mO_s^*, \qquad s > S. \tag{6.7}$$

For since $O_s \supset O_s^*$, $m(O_s\cdot\bar{O}) \geqq m(O_s^*\cdot\bar{O})$ and it follows from (6.6) and (6.7) that

$$\liminf_{s\to+\infty} m(O_s\cdot\bar{O}) \geqq \frac{m\bar{O}mO_0}{m\Omega} - \epsilon mO_0.$$

But since ϵ can be chosen arbitrarily small, (6.5) must hold.

To prove (6.7) it is sufficient to show that given $\bar{\epsilon} > 0$, there exists an $S(\bar{\epsilon})$ and for each s a measurable set $\Sigma_s^* \subset O_s^*$ such that

$$\begin{cases} m(O_s^* - \Sigma_s^*) < \bar{\epsilon}, & s > S(\bar{\epsilon}), \\ m(\Sigma_i^*\cdot\bar{O}) \geqq \dfrac{m\bar{O}}{m\Omega} m\Sigma_s^* - \bar{\epsilon} m\Sigma_s^*, & s > S(\bar{\epsilon}). \end{cases} \tag{6.8}$$

For, given ϵ, choose $\bar{\epsilon}$ such that $0 < \bar{\epsilon} < \epsilon/2$ and

$$\bar{\epsilon} < \frac{\epsilon}{3}\frac{m\Omega}{m\bar{O}} mO_0.$$

Then if $s > S(\bar{\epsilon})$,

$$m(O_s^*\cdot\bar{O}) \geqq m(\Sigma_s^*\cdot\bar{O}) \geqq \frac{m\bar{O}}{m\Omega} m\Sigma_s^* - \bar{\epsilon} m\Sigma_s^*$$

$$\geqq \frac{m\bar{O}}{m\Omega} mO_s^* - \frac{m\bar{O}}{m\Omega}\bar{\epsilon} - \bar{\epsilon}mO_s^*$$

$$\geqq \frac{m\bar{O}}{m\Omega} mO_s^* - \frac{\epsilon}{3} mO_0 - \frac{\epsilon}{2} mO_s^*, \qquad s > S(\epsilon).$$

With the aid of (6.6) there exists an $S > S(\bar{\epsilon})$ and so large that

$$\frac{\epsilon}{3} mO_0 < \frac{\epsilon}{2} mO_s^*, \qquad s > S.$$

But then

$$m(O_s^* \cdot \bar{O}) \geqq \frac{m\bar{O}}{m\Omega} mO_s^* - \epsilon mO_s^*, \qquad s > S,$$

and (6.7) holds.

We now proceed to the definition of the sets Σ_s^* such that the conditions of (6.8) are satisfied. Any point p of Ω determines, under ${}_RH_s$, a *path* in Ω, namely the set $p_s = {}_RH_s(p)$, $-\infty < s < +\infty$. This set in Ω is represented in Ψ by an infinite set of congruent right horocycles. By the *length* of the path segment $p_s p_t$ we shall mean $|t - s|$, which is also the H-length of any one of the segments of horocycles in Ψ representing this path segment.

According to Theorem 5.1 the right horocycle flow ${}_RH_s$ is metrically transitive. This implies the existence of a measurable set $\Omega^* \subset \Omega$ such that $m(\Omega - \Omega^*) = 0$, and if $p \subset \Omega^*$, the path through p is equidistributed in Ω. Specifically, if $p \subset \Omega^*$, if ν is a path segment through p and of length L, and if $\bar{L}$ denotes the linear measure of the part of ν in $\bar{O}$,

$$\lim_{L \to +\infty} \frac{\bar{L}}{L} = \frac{m\bar{O}}{m\Omega}, \tag{6.9}$$

independently of the initial point of ν, provided that it contains p (this is the ergodic theorem of Birkhoff, cf. von Neumann, p. 590, a).

Given $\epsilon > 0$ and $L > 0$, let $\Omega_\epsilon(L)$ be the set of points of Ω such that if $p \subset \Omega_\epsilon(L)$, any path segment through p and of length L has the property

$$\frac{\bar{L}}{L} > \frac{m\bar{O}}{m\Omega} - \epsilon.$$

Since this is an inequality, and the set $\bar{O}$ is an open set, the set $\Omega_\epsilon(L)$ is an open set. Let $L_1 < L_2 < \cdots$, be a sequence of positive numbers such that $\lim_{n \to \infty} L_n = +\infty$, and consider the corresponding sets $\Omega_\epsilon(L_i)$, $i = 1, 2, \cdots$. Let

$$\Omega_j = \prod_{i=j}^{\infty} \Omega_\epsilon(L_i).$$

Then $\Omega_1 \subset \Omega_2 \subset \cdots$, and since each of these sets is measurable, the same is true of the outer limiting set. Given any point $p \subset \Omega^*$, there exists an N_p such that $p \subset \Omega_\epsilon(L_i)$, $i > N_p$. Thus the outer limiting set of the sequence

$\Omega_1 \subset \Omega_2 \subset \cdots$ contains Ω^*, and we have $\lim_{j\to\infty} m\Omega_j = m\Omega^* = m\Omega$. Since $m\Omega_j \leqq m\Omega_\epsilon(L_j) \leqq m\Omega$ it follows that $\lim_{j\to\infty} m\Omega_\epsilon(L_j) = m\Omega$ and we can conclude the following. Given $\epsilon > 0$, there exists an $L(\epsilon)$ and an open set $O(\epsilon)$ such that $m(\Omega - O(\epsilon)) < \epsilon$, and if $p \subset O(\epsilon)$, any path segment of length $L(\epsilon)$ and passing through p is such that

$$\frac{\bar{L}(\epsilon)}{L(\epsilon)} > \frac{m\bar{O}}{m\Omega} - \epsilon, \tag{6.10}$$

where $\bar{L}(\epsilon)$ denotes the linear measure of the part of the segment in $\bar{O}$. Since an open set of finite measure contains a closed set whose measure differs from that of the open set by an arbitrarily small amount and since, $m(\Omega)$ being finite, the set can be chosen as compact, the following lemma can be stated.

LEMMA 6.1. *Given $\epsilon > 0$, there exists a compact, closed set $C(\epsilon)$ and a number $L(\epsilon)$ such that $m[\Omega - C(\epsilon)] < \epsilon$ and if $p \subset C(\epsilon)$, any path segment of length $L(\epsilon)$ and passing through p is such that* (6.10) *holds.*

Recall that the set O_s^* is the sum of the sets O_s^i, $i = 1, 2, \cdots, k(s)$, where each of these sets is determined by the elements on sets of arcs of right geodesic circles and of length L. A set such as O_s^i will be called a *tube* of length L, a *segment* of which is the subset determined by the elements on one of the arcs determining O_s^i. The *initial points* of the tube O_s^i are the points determined by the initial elements of the arcs determining O_s^i. The initial points of O_s^i form a two-dimensional manifold m_i which is the homeomorph of the interior of a circle. The diameter of such a tube will be defined as the diameter of its set of initial points. Given $\eta > 0$, it is evidently possible to define a set of non-overlapping tubes, each of length L, each of diameter less than η, and such that their sum is, except for a set of measure zero, identical with O_s^i. Thus the set O_s^* can be considered as the sum, except for a set of measure zero, of a set of non-overlapping tubes $\sum T_s^i$, each tube of length L, and each of diameter less than $\eta(s)$, where $\lim_{s\to\infty} \eta(s) = 0$.

Now given $\bar{\epsilon} > 0$, let $\epsilon = \bar{\epsilon}/2$, and let $L(\epsilon)$ and $C(\epsilon)$ be the length and set as determined by Lemma 6.1. Let the L used in determining O_s^* be $L(\epsilon)$. Let Σ_s^* be the subset of O_s^* formed by those tubes of the set $\sum T_s^i$ which have a point in common with $C(\epsilon)$. Then

$$m(O_s^* - \Sigma_s^*) \leqq m[\Omega - C(\epsilon)] < \epsilon < \bar{\epsilon},$$

and the first condition of (6.8) is satisfied. Suppose that the second condition of (6.8) is not satisfied. This would imply the existence of a sequence $s_1 < s_2 < \cdots$, with $\lim_{n\to\infty} s_n = +\infty$, such that

$$m(\Sigma_{s_i}^* \cdot \bar{O}) < \frac{m\bar{O}}{m\Omega} m\Sigma_{s_i}^* - \bar{\epsilon} m\Sigma_{s_i}^*, \qquad i = 1, 2, \cdots. \tag{6.11}$$

But then this inequality would necessarily hold for a tube T_i of the set forming

$\Sigma^*_{\delta_i}$, and we would have a sequence of tubes, $T_1, T_2, \cdots$, with each T_i a member of the set $\Sigma^*_{\delta_i}$, and such that

$$m(T_i \cdot \bar{O}) < \frac{m\bar{O}}{m\Omega} mT_i - \bar{\epsilon} mT_i < \frac{m\bar{O}}{m\Omega} mT_i - 2\epsilon mT_i. \tag{6.12}$$

If η_i denotes the diameter of T_i, we have $\lim_{i\to\infty} \eta_i = 0$. Each of these tubes is of length $L(\epsilon)$ and each contains a point of the set $C(\epsilon)$. The points of the tubes are determined by segments of right H-circles, the radii of which become infinite as i becomes infinite. We show that the existence of such a sequence of tubes is incompatible with the properties of $C(\epsilon)$.

To that end, let p_i be a point in T_i and in $C(\epsilon)$. Since the set $C(\epsilon)$ is closed and compact, the sequence $p_1, p_2, \cdots$, contains a subsequence which converges to a unique point p of $C(\epsilon)$. We can assume that this is true of the sequence itself or that $\lim_{i\to\infty} p_i = p$. Likewise, if q_i^α is an initial point of one of the segments of T_i, it is no restriction to assume that there exists a point q^α such that $\lim_{i\to\infty} q_i^\alpha = q^\alpha$. Given a neighborhood of q^α, since $\lim_{i\to\infty} \eta_i = 0$, for i sufficiently large, the set of initial points of the segments determining T_i lies in this neighborhood. The point q^α determines a path of the flow ${}_RH_s$. Let $q^\omega = {}_RH_{L(\epsilon)}(q^\alpha)$ and let $q^\alpha q^\omega$ be the path segment joining q^α and q^ω. As i becomes infinite, the radii of the right H-circles, segments of which determine the segments of T_i, become infinite and, again making use of $\lim_{i\to\infty} \eta_i = 0$, the segments of T_i approximate $q^\alpha q^\omega$ uniformly and arbitrarily closely. Since $\lim_{i\to\infty} p_i = p$, it follows that p must lie on $q^\alpha q^\omega$. Thus, if $\bar{L}(\epsilon)$ denotes the linear measure of the part of $q^\alpha q^\omega$ in $\bar{O}$,

$$\frac{\bar{L}(\epsilon)}{L(\epsilon)} > \frac{m\bar{O}}{m\Omega} - \epsilon.$$

Let the path segment $q^\alpha q^\omega$ be divided into ν equal closed segments of length $\delta > 0$. Let ν_δ be the number of these segments made up entirely of points u such that all points within distance δ of u lie in $\bar{O}$. It is readily shown that as δ approaches zero, the sum $\nu_\delta \delta$ of the lengths of such segments approaches $\bar{L}(\epsilon)$. Thus, by choosing $\delta = \bar{\delta}$ sufficiently small,

$$\frac{\nu_{\bar{\delta}} \bar{\delta}}{L(\epsilon)} > \frac{m\bar{O}}{m\Omega} - \frac{3}{2}\epsilon.$$

Now let the tube T_i be divided into $\bar{\nu}$ subtubes by dividing each of the segments of T_i into $\bar{\nu}$ equal segments each of length $\bar{\delta}$. As i becomes infinite each of these subtubes approaches uniformly one of the segments of length $\bar{\delta}$ of $q^\alpha q^\omega$. For i sufficiently large, $\nu_{\bar{\delta}}$ of these subtubes will lie entirely in $\bar{O}$. Since any one of these subtubes (for fixed i) can be obtained from any other one by a measure preserving right circular flow, all of these subtubes have the same measure μ_i. Thus, for i sufficiently large,

$$\frac{m(T_i \cdot \bar{O})}{mT_i} \geqq \frac{\nu_{\bar{\delta}} \mu_i}{\bar{\nu}\mu_i} = \frac{\nu_{\bar{\delta}} \bar{\delta}}{L(\epsilon)} > \frac{m\bar{O}}{m\Omega} - \frac{3}{2}\epsilon,$$

which is the desired contradiction. The second condition of (6.8) must be fulfilled by the sets Σ_i^* and the geodesic flow G_t is a mixture. We state this result as a theorem.

THEOREM 6.1. *If M is a two-dimensional manifold of constant negative curvature obtained by identifying the points of the hyperbolic plane*

$$ds = \frac{2|dz|}{1 - zz}, \qquad z\bar{z} < 1,$$

which are congruent under a Fuchsian group F which has $z\bar{z} = 1$ as principal circle, is of the first kind and has a finite set of generators, the geodesic flow on M is a mixture.

The value of the curvature of all these manifolds is -1. However, by replacing the coefficient 2 in the metric by a suitable positive constant, the curvature can be made to assume any given negative value and the proof of the stated theorem still holds. The manifolds thus obtained include, in particular, all closed, orientable, two-dimensional, Riemannian manifolds of constant negative curvature. More generally, they include the set of complete (every geodesic can be continued to infinite length) two-dimensional Riemannian manifolds of constant negative curvature, of finite connectivity and of finite (non-euclidean) area (cf. Hopf, p. 69). By including transformations of the kind

$$w = \frac{a\bar{z} + b}{c\bar{z} + d}$$

in the group F, the results hold for non-orientable manifolds. We have then the following result.

THEOREM 6.2. *The geodesic flow on any complete, two-dimensional Riemannian manifold of constant negative curvature, of finite connectivity and of finite area, is a mixture.*

BRYN MAWR COLLEGE.

BIBLIOGRAPHY

HEDLUND, G. A., *"Fuchsian Groups and transitive horocycles,"* Duke Mathematical Journal, vol. 2 (1936), pp. 530–542.

HOPF, E., *"Ergodentheorie,"* Ergebnisse der Mathematik und ihre Grenzgebiete, V2, Berlin, 1937.

MORSE, M., *"Instability and transitivity,"* Journal de Mathématiques pures et appliquées, Série 9, vol. 40 (1935), pp. 49–71.

VON NEUMANN, J., *"Zur Operatorentheorie in der klassichen Mechanik,"* Annals of Mathematics, vol. 33 (1932), pp. 587–642.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

CLASSIFICATION OF CURVES ON A TWO-DIMENSIONAL MANIFOLD UNDER A RESTRICTED SET OF CONTINUOUS DEFORMATIONS

BY C. TOMPKINS[1]

(Received September 16, 1938)

Let $\mathfrak{A}$ denote a continuous deformation of the inner equator of a torus over the surface and back onto itself in such a way that each point of the inner equator describes a path homotopic to a meridian circle of the torus. Let $l(\mathfrak{A})$ denote the maximum length to which the inner equator is stretched during the deformation $\mathfrak{A}$. The problem then arises: what is the greatest lower bound of $l(\mathfrak{A})$ as $\mathfrak{A}$ ranges over all possible deformations of the sort just described? It is shown in §8 that this greatest lower bound is precisely the length of the shortest closed geodesic on the surface which passes through a pair of opposite points on the outer equator of the torus. Furthermore, there exists a deformation $\mathfrak{A}$ for which this greatest lower bound is actually reached.

The solution obtained for this special problem in the calculus of variations in the large depends on the following "three-circle theorem":

If A, B, and C are three concentric circles in a plane and if $D(t)$, $0 \leqq t \leqq 1$, is a continuous deformation of A into C (i.e., $D(0) = A$, $D(1) = C$) then there exists a value t_0 such that the curve $D(t_0)$ intersects B in diametrically opposite points. The proof is given in §6.

This three circle theorem admits extensions and generalizations galore. One of these is presented here; the particular one chosen for consideration has been singled out because it introduces no really difficult computations and yet serves to illustrate a line of reasoning which seems to apply to many similar problems.

The proof of this extension of the three-circle theorem constitutes the main body of the paper (§§1–5), and it is the first thing to be considered. The proof of the three-circle theorem will then be presented as a corollary of the extended theorem, and finally the solution of the special problem announced above will be given.

1. Introduction and preliminary classification

Let S be a connected open region of a two-dimensional manifold with a metric; let P be a simple closed polygon of S which separates it into two disjoint components S_1 and S_2, i.e., $S_1 + S_2 = S - P$, and let there be given a topological mapping, not the identity, of P onto itself. Denote by R the space of oriented closed curves in S, and by R' the space of closed curves in S such that no curve corresponding to a point of R' intersects P in both a point and the image of this point under the given mapping. There will be presented

[1] National Research Fellow.

a natural classification of points of R' such that two points of R' may be connected with a continuous arc if, and only if, they have the same classification; this means that closed curves of S will be classified so that one can be deformed into another without the transmigrant curve at any time simultaneously intersecting a point of P and its image if, and only if, the two curves have the same classification. That the classification to be presented accomplishes this result is the principal theorem of this paper (the theorem is stated in §5).

In this section a preliminary classification of any curve C which intersects D in a finite set p of points p_i ($i = 1, 2, \cdots, m$) will be presented; a point at which C intersects P but does not actually cross it will not be counted among the points of p. In later sections loops of C will be pulled out, reducing the number of points of p, and this process will be continued until further reduction is impossible. Then an approximation to a general curve will be introduced yielding a curve which can be classified by the procedure of this section, and finally it will be shown that the classification of any curve is unique and that this classification possesses the qualities advertised above.

Before proceeding further, it is necessary to fix a suitable machinery for the classification. A definite orientation is assigned to P. A realization of the fundamental groups G_1 and G_2 of the spaces S_1 and S_2 must be adopted and retained throughout the classification. Then the fixed points of P under the transformation must be deleted from P; this will either leave P intact or split P into at most a countable number of open arcs Q_t (this family of arcs will be denoted by Q). The labelling of these arcs is retained throughout the procedure of classification.

Involved in the preliminary classification of this section are the points of p and of p', where p' denotes the set of image points of the points of p under the transformation of P. In particular p_i' will denote the image of p_i. Concerning the points of p' it is useful to know the following lemma.

LEMMA 1. *The images of points lying in the same arc Q_t all lie on one arc of Q.*

PROOF: Since the mapping is topological, Q_t is mapped either onto itself or onto the arc $P - Q_t$. In the second case, the mapping can have no more fixed points and the images of points lying in Q_t must lie in the only other arc of the family. In the first case the images of points of Q_t must lie on Q_t itself. This completes the proof of the lemma.

If C intersects P, the subscripts of the points p_i are chosen according to cyclic order of the points on C. The next step in the classification of C is to choose the finite set of arcs of Q which contain points of p or of p', and to associate with the symbol Q_t of each of these arcs an ordered set of symbols p_i and p_j' corresponding to the order of occurrence of the points corresponding to these symbols on the arc in question (this order is with respect to the order sense induced on each of the arcs of Q by the orientation assigned to P). The third step is to group subsets of the symbols which are associated with the same arc of Q into bunches. This is accomplished according to the following definition: a maximal set q of points of p all of which lie on the same arc and which are not

separated on this arc by any point of p' will be called a bunch if the images of the points of q, which (according to lemma 1) all lie on one arc of Q, are not separated on this arc by any point of p. The order relations between points of the same bunch are dropped, but the order relations between points of different bunches are retained. The final step of the classification is to associate with each point p_i an element of the fundamental group of one of the components. The method of doing this depends on whether p_i and p_{i+1} belong to the same bunch or not. If they do, the element of G assigned to p_i is just the element assigned in the realization to curves of the type of the curve constructed by following C from a point just past the point p_i to a point just before the point p_{i+1} and then returning to the start by means of an arc approximating the arc of P which lies between p_i and p_{i+1} and which contains no points of p'. If p_i and p_{i+1} do not belong to the same bunch the element assigned to p_i is that corresponding to curves of the type of the curve obtained by passing along C from a point close to p_i to a point close to p_{i+1} as before and returning along an arc approximating the arc of P which joins p_i to p_{i+1} (as opposed to the arc which joins p_{i+1} to p_i ; this distinction again arises from the orientation assigned to P).

Two curves C and C' will be called equivalent under this preliminary classification if they intersect P the same number m of times and (i) if $m \neq 0$, there exists a cyclic permutation of the numbers $1, 2, 3, \cdots, m$ so that when it is applied to the indices of the intersections of C' the formal classification is exactly the same as that of C; or (ii) if $m = 0$, either the same element of the fundamental group G_1 or G_2 is assigned to curves of their types or else each corresponds to the identity element but in different components.

LEMMA 2. *If C_1 and C_2 are equivalent under this classification, there exists a homotopic deformation of C_1 into C_2 using only curves corresponding to points of R'.*

PROOF: If the number of intersections of C_1 with P is zero, the lemma is trivial. If there are intersections, it is evident that each of these intersections can be made to coincide with the similar one of C_2 by means of a deformation of the type allowed. Indeed, suppose that the intersections of C_1 with P are denoted by $p_{1,i}$ and those of C_2 with P by $p_{2,i}$ with the indices chosen so that the equivalence in classification is apparent with no further change of index. Some points $p_{1,i}$ may coincide with the corresponding points $p_{2,i}$, and these will be left alone for the time. Any other point $p_{1,i}$ which does not coincide with the corresponding point of the other curve may be deformed into such a point by means of a homotopy on the allowed curves which moves points between its first and last positions and points corresponding to these by any iteration of the mapping, and only those points, and which does not change the classification of the curve. Now, the points which are forcefully moved can not be points which are already lying on their corresponding points on the other curve, or the order relations between points of intersection of the two curves of P would not be the same, and they would not have the same classification. The points of $p_{1,i}$ can thus be made one at a time to coincide with

those of $p_{2,i}$ until the two sets coincide exactly. After this the arcs of the transform of C_1 may be deformed in the components of S_1 and S_2 without again crossing P until the two curves coincide throughout.

2. Reduction of a class

There will be described here a means of deriving, from a class of curves with m intersections with P, a reduced class which is either the same or has fewer intersections with P. The points to be removed are maximal sets of p each containing more than one point such that (i) all the points of the set belong to the same bunch; (ii) those in each bunch have consecutive indices; and (iii) all but the last in each set have associated with them identity elements of the groups G_1 and G_2. If the number of elements in a set which is removed is odd the set is replaced by one point having associated with it the order relations of the bunch from which the set is removed and the group element of the last point of the set removed. If the number is even, nothing replaces the set removed, but the group element associated with the point which immediately precedes the set is replaced by the group element obtained by multiplying it on the right by the element associated with the last point of the removed set and then, if the point immediately following the removed set is in the same bunch as the one which immediately precedes it, but not in the same bunch as the points of the set, or if the point immediately following the set lies between the point preceding the set and the set itself, this product is again multiplied on the right by the element of the fundamental group in which the first product lies and which corresponds to a curve approximating P and having the same orientation. The images of the deleted points are also deleted, the residues of the sets p and p' are grouped into a new set of bunches and the group elements associated with points of new bunches are changed by multiplying by P as above if necessary, and the new class is the result. If all points of p are removed in this process the new classification is represented by just the proper group element.

LEMMA 3. *There exists a homotopic deformation of a curve C of class $\mathfrak{a}$ into a curve C' of the class $\mathfrak{a}'$, the reduced class of $\mathfrak{a}$.*

PROOF: Let C be any curve of class $\mathfrak{a}$, and deform it homotopically by moving arcs corresponding to those lost in the reduction keeping them always on their original components S_1 and S_2 until they approach arbitrarily closely arcs of P none of which contain points of p'. They may then be deformed onto these arcs and then removed in any way onto $S - P$ so as to achieve a curve of the class $\mathfrak{a}'$. This curve may then be deformed into any other curve of this class because of lemma 2.

3. Complete reduction of a class

A class $\mathfrak{a}^*$ is said to be the complete reduction of the class $\mathfrak{a}$ if the reduction process when iteratively applied to $\mathfrak{a}$ until the resulting class is invariant under

the operation yields the class $\mathfrak{a}^*$. As an immediate consequence lemma 4 follows.

LEMMA 4. *There exists a homotopic deformation on curves of R' of any curve C of class $\mathfrak{a}$ into a curve C^* of class $\mathfrak{a}^*$, the complete reduction of $\mathfrak{a}$.*

4. Classification of general curves of R'

The preliminary reduction of the most general curve of R' is not unique but rather depends upon a more or less arbitrary polygonal approximation to the given curve. Let C be any curve of R'; the set D of intersections of C with P is closed, and the set D' of images of points of D under the mapping of P onto itself is also closed. Since C belongs to R', D and D' do not intersect, so there is a positive number b such that if p and q are two points of P whose distance is less than b the distance between their images p' and q' under the mapping of P is less than $a/2$. Construct an open non-singular band B of S around P and choose a positive number c smaller than each of the numbers $a/2$ and b and smaller than the distance from P to $S - B$. Finally, choose a number d as small as c and so small that every curve with Fréchet distance[2] less than d from C is homotopic with C. Let E represent the closed set of point s of C which are not joined with some point of D by an arc of C lying wholly in a d-neighborhood of D, and let the positive number e denote the distance between the disjoint closed sets D and E; if the set E is vacuous, simply write $e = d$. Now make a polygonal e-approximation C_1 on S to C so that C_1 intersects P only a finite number of times. This is the preliminary reduction of C. The classification of C is achieved by obtaining the complete reduction of the classification of C_1.

5. Uniqueness of classification; principal theorem

The nature of the first reduction of general curves of R', the reduction based on inequalities throughout, immediately yields the following lemma.

LEMMA 5. *If C is any curve of R' there exists a number f such that if C' is any other curve of R' with Fréchet distance less than f from C, C and C' admit the same classification.*

The rôle of the inequalities is to permit the same approximating polygon to be substituted for both C and C'.

The final preliminary result is stated in the next lemma, which will be proved by induction.

LEMMA 6. *The classification of any curve C is unique.*

PROOF: The only place where the procedure of classification is not uniquely determined is in the choice of the approximating polygon C_1. Now suppose that C_1 and Γ_1 are two allowed approximations of C. The degree of approximation is so chosen that the only difference between the first classification of

[2] In this case, this means that the two curves can be parameterized simultaneously so that points corresponding to the same parameter are within a distance d of each other.

these curves is in the existence of lemeents in C_1 not in Γ_1 which are of the type which disappear in the passage to C_2, and the presence of similar elements in Γ_1 which are not in C_1. (These are elements corresponding to arcs of C_1 and Γ_1 with which the identity element is associated and which subtend no point of the point set D' or of the set Δ' and which subtend no fixed point of the mapping of P.) It will be shown that the addition of any number of arcs of this type to a curve does not affect the final classification. Indeed, consider any curve with a finite number of intersections with P and add a finite number of arcs of this type. In the first reduction of the preliminary classification of this curve all these arcs are removed and possibly some arcs which would be removed from the original curve are also removed. Thus the difference between the momentary classification of curves reduced one step lies in the presence of extra arcs on one curve which will be removed in the next step, exactly the difference at the start. Thus this difference persists through each step of the reduction, and after the final step of the reduction of the simpler curve one more reduction leaves this classification invariant and causes the classification of the other to agree with that of the first. Then, returning to the lemma, the equivalence of C_1 and Γ_1 is achieved by starting with a curve C_1 which is the obvious common part of C_1 and Γ_1 and adding arcs to it, first to get C_1, and then to get Γ_1, never changing the classification.

THEOREM. *A necessary and sufficient condition for the existence of a continuous deformation in R' of a curve A into a curve B is that the classification of A and B are the same.*

The sufficiency is a consequence of lemma 4 and the accuracy of the approximation used in the classification.

If the condition is not necessary there exist curves A and B of R' and a deformation on curves of R' of A into B, such that the classifications of A and B differ. Suppose this deformation is $C(t)$, $0 \leqq t \leqq 1$, with $C(0) = A$, $C(1) = B$. Then from lemma 5, curves with t near zero admit the same classification as A; let g be the greatest number such that all $C(t)$ with $t < g$ admit this classification. If $C(g)$ admits this classification then curves with slightly higher values of t also admit it, by lemma 5, so g is not the highest number as demanded. On the other hand if $C(g)$ admits any other classification so do the curves $C(t)$ with t neighboring and below g. This contradicts lemma 6 and proves the theorem.

6. The three circle theorem

The three circle theorem of section 1 will now be established as a corollary of the main theorem. S will be taken as the plane, P will be the circle B of lemma 1, the mapping will be the mapping which makes a point of B correspond to the diametrically opposite point. Then the classification of the circle A is expressed just by the identity element, the classification of C is expressed by a generating element of the fundamental group of the region outside the circle B, which is certainly not the identity element. Then the main theorem yields the three circle theorem as a direct consequence.

7. The special problem

The special example concerns deformations on a torus whose explicit parametric representation may be given

$$x = (a - b\cos v)\cos u,$$

$$y = (a - b\cos v)\sin u,$$

$$z = b\sin v.$$

In the space R_1 of closed curves on this torus the equation $v = 2\pi t$ $(0 \leqq t \leqq 1)$ represents an arc with parameter t (i.e., to each value of t there corresponds a closed curve on the torus). This arc begins and ends with the point corresponding to the inner equator of the torus. Arcs which in R_1 are homotopic with this arc will be admitted and with any such arc $\mathfrak{A}$ a number $l(\mathfrak{A})$ is introduced. This number, positive and possibly infinite is the length of the longest curve on the torus corresponding to a point of $\mathfrak{A}$. There will be asked two questions:

(i) What is the greatest lower bound of $l(\mathfrak{A})$ for all admissible arcs $\mathfrak{A}$?

(ii) Does there exist a deformation $\mathfrak{B}$ of the admitted class such that $l(\mathfrak{B})$ is this greatest lower bound?

Before the actual answer to these questions is obtained the following lemma will be proved.

LEMMA 7. *If $\mathfrak{A}(t)$, $0 \leqq t \leqq 1$, is any allowed deformation of the inner equator of the torus around the torus and back onto itself, there exists a value t_0 of t such that the curve $\mathfrak{A}(t_0)$ intersects the outer equator in diametrically opposite points.*

PROOF: Map the torus on to a plane with polar coördinates $r = e^v$, $\vartheta = u$. Then the deformations in question are mapped into continuous deformations of the curve $r = 1$ into the curve $r = e^{2\pi}$, and the curve $r = e^{\pi}$ corresponds to the outer equator. Then, from the three circle theorem, at some stage of the deformation the image of the transmigrant curve cuts the image of the outer equator in diametrically opposite points and so, since the mapping is a one to many mapping, that is, since the mapping of the torus on the plane has a unique inverse, the transmigrant curve must intersect the outer equator in diametrically opposite points.

8. The deformation

In this section a knowledge of Bliss's paper on geodesics on a torus[3] is necessary. There are four paths of minimum length connecting a pair of diametrically opposite points of the outer equator. Suppose $v = f(u)$, $0 \leqq u \leqq \pi$, $0 \leqq v \leqq \pi$, is one of these, with $f(0) = f(\pi) = \pi$. The function $f(u)$ has one minimum in this range at $u = \pi/2$. At all other points its derivative differs from zero. Suppose $f(\pi/2) = c$. There will be constructed here a deformation

[3] Gilbert Ames Bliss, *The geodesic lines on the anchor ring*, Annals of Math., series 2, vol. 4 (1902), pp. 1–21.

of the type sought of which the longest curve will be the curve obtained by adding to this curve the similar curve with $\pi \leqq u \leqq 2\pi$ and $\pi \leqq v \leqq 2\pi$. (The curve so obtained is, because of symmetry, a closed geodesic.) It follows from lemma 7 that a curve of at least this length must be included in any allowed deformation, and so this deformation completes the answer to both the questions of §7.

(i) *The greatest lower bound of the numbers $l(\mathfrak{A})$ is the length of the shortest closed geodesic (of the topological type of the curves of $\mathfrak{A}$) which passes through diametrically opposite points of the outer equator;*

(ii) *This bound is an actual minimum which can be realized.*

The actual deformation is defined as follows. Suppose the closed geodesic already defined is denoted by the equation $v = f(u)$, $-\infty < u < \infty$. Then the deformation is

for $0 \leqq t \leqq 1/4$ and $-\pi/2 \leqq u \leqq 3\pi/2$, $v = 4ct$;
for $1/4 \leqq t \leqq 1/2$ and $|u - \pi/2| \leqq -4\pi t + 2\pi$, $v = 4(\pi - c)t + 2c - \pi$;
for $1/4 \leqq t \leqq 1/2$ and $3\pi/2 \geqq u \geqq -4\pi t + 5\pi/2$, $v = f(u + 4\pi t - 5\pi/2)$;
for $1/4 \leqq t \leqq 1/2$ and $-\pi/2 \geqq u \geqq 4\pi t - 5\pi/2$, $v = f(u - 4\pi t + 5\pi/2)$.

This completes the definition of the deformation for $0 \leqq t \leqq 1/2$. For $1/2 \leqq t \leqq 1$ the analogous deformation with center at $(3\pi/2, 2\pi - c)$ instead of $(\pi/2, c)$ is used. That the lengths of the curves depend in a monotonically increasing fashion on t for $0 \leqq t \leqq 1/2$ is obvious from a direct calculation (very simple if accompanied by a diagram). The curve for $t = 1/2$ is just the geodesic in question. This completes the demonstration.

It might be remarked here that this property of the closed geodesic is related to known properties of closed geodesics. The questions (i) and (ii) of §7 were posed by Professor Marston Morse, and he gave a correct guess as to the answer to question (i); everyone would expect the affirmative answer to question (ii).[4]

THE INSTITUTE FOR ADVANCED STUDY.

[4] H. Pétard, Am. Math. Monthly, vol. 45 (1938), pp. 446–7.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

FRÉCHET DEFORMATIONS AND HOMOTOPY

BY C. TOMPKINS[1]

(Received September 12, 1938)

1. Introduction

In the calculus of variations the type of continuous deformation of a curve which arises naturally is a deformation continuous in the sense of Fréchet distance (see the next paragraph for definitions), whereas in the classical topological treatment of continuous deformations an entirely different definition of continuous deformation is adopted. The difference lies in the distinction between parameterized curves—those appearing in the homotopy theory of topology—and unparameterized curves, which seem naturally suited to the calculus of variations. To apply a Fréchet metric to parameterized curves is unreasonable, for then distinct curves have a distance zero from each other, but it is clear that if this is done—or if a homotopic deformation is taken with its curves considered as unparameterized—every homotopic deformation is a Fréchet deformation. Something close to the converse of this statement may be obtained in spaces which are sufficiently well-behaved that in a fairly uniform way every two curves with sufficiently small Fréchet distance belong to the same homotopy class; in this case a Fréchet deformation can be replaced in an obvious way by a homotopic deformation with the same first and last curves. The principal theorem of this paper is a little more accurate than this in that it guarantees the possibility of simultaneous parameterization of all curves in a Fréchet deformation in a strongly continuous manner, thus substituting for the Fréchet deformation in any metric space a homotopic deformation.

THEOREM. *Let* $C(t)$, $0 \leqq t \leqq 1$, *represent a Fréchet deformation of a closed curve in a metric space. Then for each value of* t *the curve* $C(t)$ *may be parameterized with a parameter* u *in such a way that the curve is expressed as a point function* $p(t\,;u)$ *which is continuous in both variables* t *and* u.

This theorem has been proved for arcs by M. Morse, and the demonstration here will depend primarily upon Morse's result and a paper by H. Whitney, whose results also figure in Morse's demonstration.[2] The technique here will involve the substitution of a Fréchet family of arcs for the family of closed curves and the application of Morse's results.

---

[1] National Research Fellow.

[2] These results are in the following papers:

M. Morse, *A special parameterization of curves*, Bulletin American Mathematical Society, volume 42 (1936) pp. 915–922.

H. Whitney, *Regular families of curves*, Annals of Mathematics, second series, volume 34 (1933) pp. 244–270. These papers will be referred to as M and W respectively.

As to notation: $\rho(a, b)$ represents the distance between a and b, where a and b may be points or sets of points. (C_1, C_2) represents the Fréchet distance between curves C_1 and C_2. $W_\epsilon(A)$ is the set of all points p such that $\rho(p, A) \leqq \epsilon$.

2. Fréchet Deformations

The Fréchet distance $[C_1, C_2]$ between two sensed closed curves (continuous images of a circle) C_1 and C_2 will be defined as follows. Suppose C_1 is given as a continuous point function defined on a circle C and C_2 is given as another point function defined on the same circle. Consider any two sense preserving homeomorphisms H_1 and H_2 of C onto itself. Then a new parameterization is introduced by H_1 on C_1 and by H_2 on C_2. Let $d(H_1, H_2)$ represent the least upper bound of distance between pairs of points, one on C_1 and one on C_2, corresponding to the same values of the parameter of C after the homeomorphisms. The greatest lower bound of the numbers $d(H_1, H_2)$ for all allowed homeomorphisms is the Fréchet distance (C_1, C_2).

DEFINITION. *A one parameter family of sensed closed curves $C(t)$, $0 \leqq t \leqq 1$, is a Fréchet deformation if for every positive number ϵ there exists a positive number δ such that $[C(t_1), C(t_2)] < \epsilon$ whenever $| t_1 - t_2 | < \delta$.*

The chief result of this paper will be that it is possible to introduce a parameter u simultaneously on all the curves $C(t)$ of a Fréchet deformation so that each curve will be given as a point function $C(t ; u)$ continuous in both the variables. In this sense a Fréchet deformation is equivalent to a homotopy. This is the result Morse proved for arcs.

3. The function μ

The principal tool used here is Whitney's function $\mu(A)$ whose argument is any point set of a separable metric space R. To define this function choose a sequence of points dense over R, $\{a_i\}$.
Let p be any point of R and define

$$f_i(p) = \frac{\rho(p, a_i)}{1 + \rho(p, a_i)}.$$

Then let A be any point set of R and define

$$\mu_i(A) = \text{L.U.B. } f_i(p) - \text{G.L.B. } f_i(q), \qquad p, q \in A$$

and finally,

$$\mu(A) = \sum \mu_i(A)/2^i$$

The properties of the function $\mu(A)$ which make it important are principally these:

LEMMA 1. *If $A \subset W_\epsilon(A')$, with $\epsilon > 0$, then*

$$\mu(A) < \mu(A') + 2\epsilon \qquad \text{(W. p. 246)}$$

LEMMA 2. *If $A \subset A'$ and A' contains a point at positive distance from A, then*

$$\mu(A) < \mu(A') \qquad \text{(W. p. 246)}$$

LEMMA 3. *If A is any closed point set, (possibly vacuous) and q_0q_1 a simple arc with an interior point p not in A, and if $\mu(A + q_0p) = \alpha$, then for any positive number ϵ there exists a positive number δ such that if $q_0'q_1'$ is an arc with Fréchet distance from $q_0q_1[q_0q_1, q_0'q_1'] < \delta$ and if p' is a point of $q_0'q_1'$ such that $\mu(A + q_0'p') = \alpha$, then $\rho(pp') < \epsilon$.* (W. pp. 256–259.)

PROOF OF LEMMA 3. Suppose the proposition false; then there exists a point set A, an arc q_0q_1 and a sequence of arcs $\{q_0^iq_1^i\}$ approaching q_0q_1 in the sense of Fréchet and there exist points p on q_0q_1 and p^i on $q_0^iq_1^i$ such that $\mu[A + pq_0]$ and $\mu[A + p^iq_0^i]$ are both equal to α and such that the distance from p to any p^i is at least ϵ. It is clear that there is a subsequence of points $\{p^i\}$ which converges to a limit point p^* on q_0q_1, and this point has a distance at least ϵ from p. Two cases arise: p^* lies between p and q_0 or between p and q_1; the treatment of the second case involves interchanging the rôles of p and p^* in the treatment of the first, which will be presented here.

Choose $\sigma > 0$ so that $3\sigma < \rho(p, p^*q_0)$ and $3\sigma < \rho(p, A)$, and choose a base point a_i of the set of points on which the definition of the function μ depends so that $\rho(a_i, p) < \sigma$. Then $\rho(a_i, p^*q_0) > 2\sigma$ and $\rho(a_i, A) > 2\sigma$; from this it follows that $\mu_i(pq_0 + A) > \mu_i(p^*q_0 + A) + \sigma$ for the particular choice of index i while for all other indices $\mu_n(pq_0 + A) \geqq \mu_n(p^*q_0 + A)$. From this it follows that $\mu(pq_0 + A) > \mu(p^*q_0 + A) + \sigma/2^i$. But there exist arcs of the sequence $\{q_0^nq_1^n\}$ such that $p^nq_0^n$ is as close as is wanted to p^*q_0; in particular this distance may be taken to be less than $\sigma/2^{i+1}$. Then by lemma 1 $\mu(p^nq_0^n + A) < \mu(p^*q_0 + A) + \sigma/2^i$. The left member of this inequality $\mu(p^nq_0^n + A) = \alpha$. Combining it with the preceding inequality the contradiction $\alpha < \mu(p^*q_0 + A) + \sigma/2^i < \mu(pq_0 + A) = \alpha$ is obtained.

4. The associated family and its expansion

Let S be a metric space, S' the space of sensed arcs in S metrized with the Fréchet metric and finally let C be a sensed closed curve in S. To any point p on C there corresponds an arc in S obtained by starting at p and tracing out C in a positive sense, ending again at p: this point is represented by a point in the space S'. Suppose that C is represented as a continuous function of a point u on a circle; then this parameterization is induced on the set of points in S' associated with the choices of p. Under this parameterization it will be shown that this set is a closed curve: the points in S' are also continuous functions of the parameter u. In anticipation of this result this set of points considered as an unparameterized point set with order relations agreeing with those of C will be called the curve associated with C.

LEMMA 4. *The curve D associated with a closed curve C is the continuous image of a circle.*

PROOF. Represent C as a continuous point function $p(u)$ of a parameter u;

since u is defined over a closed point set (a circle) the function $p(u)$ is uniformly continuous. Then for any positive number ϵ there exists a positive δ such that $\rho[p(u_1), p(u_2)] < \epsilon$ whenever $| u_1 - u_2 | < \delta$. It will be shown that if $| u_1 - u_2 | < \delta$ the points in S' corresponding to the arcs of $p(u_1)$ and $p(u_2)$ are within a distance less than ϵ of each other. Indeed leave the parameter u intact on the first of these arcs and substitute the parameter $\bar{u} = u + u_1 - u_2$ on the second. Then the continuity inequality states that under this parameterization corresponding points lie within a distance ϵ of each other, and this implies that the Fréchet distance between the arcs is less than ϵ, which was chosen arbitrarily.

LEMMA 5. *Let $C(t)$, $0 \leqq t \leqq 1$, be a Fréchet deformation in S and let $D(t)$ represent the family of associated curves in S'; then $D(t)$ is a Fréchet deformation.*

The proof will follow from a demonstration that $(D(t_1), D(t_2)) = (C(t_1), C(t_2))$, which is immediately evident.

Other properties of the Fréchet deformations are stated.

LEMMA 6. *If $C(t)$ is a Fréchet deformation and $\{t_i\}$ a sequence of values of t approaching $\bar{t}$, and if $p(t_i)$ represents a point on $C(t_i)$ for each index i, the sequence $\{p(t_i)\}$ has at least one limit point on $C(\bar{t})$.*

LEMMA 7. *If $C(t)$ is a Fréchet deformation the set of points contained in the sum of all curves $C(t)$ is connected.*[3]

LEMMA 8. *The curve D of S' associated with a curve C of S is either a simple closed curve or a simple closed curve taken several times.*

PROOF. Let C be given a parameterization; then it is given as a uniformly continuous image of a circle, B. This induces a parameterization on D. The proof follows from the italicized statements below, each of which is proved. *If p and q are any points of B which are in different components of the inverse image of one point P of D, then the diameter of C is the diameter of the image of the arc pq.* Since p and q correspond to the same point on D, they correspond to two coincident arcs on C with both end points of each arc at one point, the image under the first parameterization of both points p and q. Thus the arc pq goes into a closed curve, which is the beginning of the arc corresponding to say p. At q the arc corresponding to q begins, and it must coincide with the arc starting with p, so it must run through this same closed curve, which now is a second section of the arc of p; the arc of q must therefore follow it again, and an inductive argument shows that *the curve C simply traces this curve a finite number of times.* The next statement: *If C is not a single point the distances between any components of the inverse image of a point P of D are bounded away from zero uniformly for all P.* Otherwise there would be two points of different components arbitrarily close to each other, the uniform continuity of C would, with the first italicized statement make the diameter of C arbitrarily small, a contradiction. An immediate corollary is that *the inverse image of a point of D contains only a finite number of components, and a neighborhood of this inverse image on B corresponds to a finite number of simple arcs through the point in question on D.*

[3] See R. L. Moore, *Foundations of point set theory*, New York (1932) p. 19, Theorem 32.

Now if D is not a simple curve covered once choose a point P of D and two successive components of the inverse image of P on B. If the arc joining these components is not simply mapped on D (under a simple mapping it is allowable for an arc to go into a point—a proper name would be a monotone mapping) this process may be repeated on this arc. Due to uniformity conditions stated above a finite number of iterations of this process yields an arc of B mapped simply onto a closed curve of D, and the second italicized statement completes the proof.

The final important property of the family adjoined to a Fréchet family has to do with its regularity in the sense of Whitney (W. p. 248). A family of curves F is defined by Whitney to be regular if (1) no two curves of the family intersect, and (2) given any arc pq of a curve C of the family and any positive number ϵ, there exists a positive number δ such that if p' is a point with distance less than δ from p lying on a curve C' of the family, there is an arc $p'q'$ of C' such that $(pq, p'q') < \epsilon$. The second condition obviously guarantees the property announced in lemma 8 for any curve D of a regular family.

LEMMA 9. *If $C(t)$ is a Fréchet deformation and $D(t)$ the associated deformation then $D(t)$ satisfies the second regularity condition.*

PROOF. Choose the arc pq of a curve $D(t_1)$ and choose a positive number ϵ. Now suppose a point p' of a curve $D(t_2)$ (t_2 not necessarily different from t_1) has a distance less than ϵ itself from p; then there exist simultaneous parameterizations of $C(t_1)$ and $C(t_2)$ under which p and p' correspond to the same values of the parameter and such that no two corresponding points of the two curves have a distance from each other as great as ϵ. This parameterization is induced on $D(t_1)$ and $D(t_2)$; let q' be the point on $D(t_2)$ corresponding to the same value of the parameter as q. Then obviously $(pq, p'q') < \epsilon$.

In order to achieve the first regularity condition the family $D(t)$ will be replaced by a family $D_1(t)$ obtained by expanding $D(t)$ along an extra dimension added to the space S'. Specifically, introduce the space S_1 which is the product space of S' with the line interval $0 \leqq t \leqq 1$ and metrize S_1 so that the distance between two points is the distance between their images in S' plus the absolute value of the difference of their t-coördinates. For any curve $D(t_1)$ of S' substitute the curve $D_1(t_1)$ of S_1 whose image in S' is just $D(t_1)$ and which has the value t_1 for the t-coördinate of each of these points. In this fashion the family $D_1(t)$ $0 \leqq t \leqq 1$ is built up. The first regularity condition is satisfied: no two curves of the family $D_1(t)$ intersect. It is easily seen that all the lemmas of this section, lemmas four through nine, are valid if this family $D_1(t)$ is substituted for the family $D(t)$, and in addition the following statement is true.

LEMMA 6A. *If $\{t_i\}$ is a sequence of numbers $0 \leqq t_i \leqq 1$ and $p(t_i)$ is a point of $D_1(t_i)$ for each index i, then if the sequence $\{p(t_i)\}$ has a limit point on a curve $D_1(\bar{t})$, $\bar{t}$ must be a limit value of the sequence $\{t_i\}$.*

In this section there has been substituted for a Fréchet deformation in S the expanded deformation associated with it in the space S_1—the product space of the space of arcs of S with a straight line segment—which is equivalent to the

original deformation and which has the property of being a regular family of curves in the sense of Whitney.

Now the space S_1 will be discarded and of it only the family $D_1(t)$ retained. This must be done in order to have a separable metric space, the space composed of points on $D_1(t)$, in which to set up the function μ.

LEMMA 10. *The space of points of $D_1(t)$ is separable.*

PROOF. Each curve of $D_1(t)$ is separable, for each curve is a continuous image of a circle. Then for a countable dense set of values of the parameter t, $\{t_i\}$ on each of the curves $D_1(t_i)$ may be chosen a countable dense set of points. The complete set of points so obtained is necessarily countable, and that it is dense follows from the continuity of the family $D_1(t)$. Indeed if p is any point of $D_1(t)$ and d any positive number, infinitely many points of the set within a distance d of p may be found as follows. Choose a curve of $\{D_1(t_i)\}$ with Fréchet distance less than $d/2$ from the curve of p; on this curve choose a point p' whose distance from p is less than $d/2$; also on this curve are infinitely many points of the countable dense set with distance less than $d/2$ from p', hence with distance less than d from p.

5. Cross sections

If p is any (interior) point of a curve C of a regular family F, Whitney (W. p. 256) has defined a cross section through p. A set of points S form a *cross section through p* if the following conditions are satisfied:

(1) Each point p' of S lies within an arc $q_0'q_1'$ of one of the curves of F such that for some $\lambda' > 0$ each arc of F with Fréchet distance from $q_0'q_1'$ less than λ' contains at most one point of S.

(2) p lies within an arbitrarily small arc q_0q_1 of a curve of F such that for some $\lambda > 0$ each arc of F with Fréchet distance from q_0q_1 less than λ contains exactly one point of S.

(3) S is closed.

Since Whitney was presumably considering simple curves the following condition must be added to his when curves which are not necessarily simple are considered.

(4) If a point p' of S lies on two arcs of any curve it is considered as belonging to each of them.

Continuing to follow Whitney almost verbatim, let F be a regular family the diameters of whose curves are bounded away from zero, let C be a curve of F, q_0q_1 be an arc of C and p an interior point of q_0q_1 (q_0q_1 is to be a simple arc). Choose σ so that $4\sigma < \rho(q_0, pq_1)$ and $4\sigma < \rho(q_1, pq_0)$ and define $U = W_{2\sigma}(q_0)$. If any curve C^* of F contains a point p^* near p, call $q_\zeta (\zeta > 0)$ the first point of C^* in $W_\zeta(q_0)$ (if it exists) following C^* from p^* towards U; call q_ζ the first point of C in W_ζ. Choose η so small that the arc $q_{2\eta}q_0 \subset V_\sigma(q_0)$ and define $\mu(U + pq_\eta) = \alpha$.

THEOREM 1. *On each arc near q_0q_1 there is a point p' for which $\mu(U + p'q_\eta') = \alpha$ (and these points form a cross-section through p).*

The proof of this theorem is a reasonably accurate facsimile of Whitney's proof (W. pp. 257–259), and so is omitted. The only addition to this proof necessary is the remark that condition (4) of a cross section is obviously satisfied. What is important here is the existence of p', which follows from lemma 3.

THEOREM 2. *Let $D_1(t)$ be the expanded family associated with a Fréchet deformation $C(t)$ and suppose the diameters of curves of $D_1(t)$ are bounded away from zero. Then there exists near any point p of a curve $D_1(t_0)$ a set of points $p(t)$, t near t_0, such that $p(t)$ lies on $D_1(t)$ for each t, $p(t)$ depends continuously on t, and $p(t_0) = p$.*

PROOF. Construct a cross section $S(p)$ through p by the method outlined above. If for t near t_0 $S(p)$ contains only one point on each curve $D_1(t)$ the theorem follows immediately from lemma 3. This includes the case where a point of S lying on several branches of $D_1(t)$ is counted as one point.

If some curves near $D_1(t_0)$ contain several points of S, recourse to lemma 3 again yields the result that S contains an arc through any point of S with the property that near this point there is only one point of S on each curve of $D_1(t)$. The set of points of S which do not satisfy this condition is countable, for the condition is satisfied for points on all curves except those which are simple closed curves taken several times and which are limit curves of simple closed curves taken fewer times; these curves will be called *exceptional curves*. The exceptional values of t, that is the values of t such that $D_1(t_i)$ are exceptional curves, form a closed discrete point set. Now it is clear that an arc of S through a point near p can be continued to the first exceptional curve on each side of the point, and that along with a point on each of these exceptional curves this extended arc forms a closed arc. This last is true because the arc cannot approach more than one point on the exceptional curve without violating the first conditions for cross sections. In this fashion arcs on each of the open intervals of non-exceptional values of t may be constructed and end points added from exceptional curves. Then in the usual fashion a countable number of these arcs may be joined to form an open arc which must approach just one point, the point p, on $D_1(t_0)$. This family of arcs then yields an arc through p.

THEOREM 3. *An expanded family $D_1(t)$, $0 \leqq t \leqq 1$, associated with a Fréchet deformation $C(t)$ admits a set of points $p(t)$, $0 \leqq t \leqq 1$, such that for each t, $p(t)$ lies on $D_1(t)$ and such that $p(t)$ is a single valued continuous point function of t.*

PROOF. Suppose first that the diameters of curves of $D_1(t)$ are bounded away from zero. Let p be a point on $D_1(0)$; then a function $p(t)$ for t near zero may be constructed such that $p(0) = p$ (Theorem 2). Suppose this function can not be extended all the way across the range of t values. Then there is a least upper bound of t values for which the function may be defined. Call this value a, $0 < a \leqq 1$. If the function $p(t)$ approaches a point $p(a)$ on $D_1(a)$ it can be extended by adding an arc through $p(a)$, so $p(t)$ must converge to an arc of $D_1(a)$. But in this case, if $a < 1$, an arc through any interior point of this convergence arc constructed as in theorem 2 will intersect $p(t)$ in an interior point. If the later part of $p(t)$ is replaced by this arc $p(t)$ will be extended

past the value $t = a$. Finally, this scheme may be used if $a = 1$ to obtain a function $p(t)$ $0 \leqq t < 1$ which converges to just one point on $D_1(1)$, thus completing the definition of $p(t)$ for the range $0 \leqq t \leqq 1$.

Now if there are curves of zero diameter, each of these curves consists of just one point. The values t_i such that $D_1(t_i)$ are curves with diameter zero and such that in every neighborhood of t_i there are curves with non-zero diameter form a discrete closed set. It is evident that a countable family of arcs drawn toward one of these points converges to the point, and this permits extension of an arc completely across the family.

COROLLARY. *Let $D(t)$ be the family associated with a Fréchet deformation $C(t)$, $0 \leqq t \leqq 1$. There exists a continuous single valued point function $p(t)$ with $p(t)$ lying on $D(t)$.*

This function is the image of the function constructed in theorem 3.

The principal theorem of the paper is next.

THEOREM 4. *Let $C(t)$, $0 \leqq t \leqq 1$, be a Fréchet deformation of a closed curve in a metric space. Then for each value of t the curve $C(t)$ may be parameterized with a parameter u in such a way that the curve is expressed as a point function $p(t\,;u)$ which is continuous in both variables t and u.*

PROOF. The result of the preceding corollary is that deformations of closed curves may be regarded as Fréchet deformations of closed arcs. The continuous parameterization of these arcs has been given by Morse (M. p. 921), and the parameterization, considered as a parameterization of the closed curves, yields the desired result.

UNIVERSITY OF MICHIGAN.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

THE GROUP OF ISOMETRIES OF A RIEMANNIAN MANIFOLD

BY S. B. MYERS AND N. E. STEENROD

(Received February 23, 1938; Revised November 7, 1938)

1. Introduction

In the classical theory of groups of motions (isometries) of a Riemannian space,[1] neither "whole" groups nor "whole" spaces were ever considered, but only "group germs" and spaces in the neighborhood of a point. Furthermore one assumed given a *Lie group germ* of motions and then deduced theorems concerning the structure of the space. The present paper contains a study of the total group of isometries of a Riemannian manifold in the large. The manifold M is taken of class C^r $(r \geqq 2)$; and by an isometry is meant any distance preserving transformation. A natural topology is defined in the group G of all isometries. The motion of a point under the group is studied and is found to describe a submanifold of class C^1. This leads to the introduction of parameters in G from which it follows that G is locally euclidean. It is then proved that, under these parameters, *G is a Lie group of transformations of M.* The same proposition is proved for any closed subgroup of G. Thus the classical theory of Lie groups of motions applies to any closed group of motions.

2. Lie Groups

An r-parameter Lie group is a topological group whose group space is an r-dimensional manifold in which coördinate neighborhoods of class C^1 (see §3) have been introduced in such a way that the law of composition $C = B \cdot A$ is described locally by functions $\gamma^i = \phi^i(\alpha^1, \cdots, \alpha^r; \beta^1, \cdots, \beta^r)$ of class C^1 ((α), (β), (γ) being coördinates for neighborhoods of A, B, C respectively).[2]

An r-parameter Lie group of transformations is an r-parameter Lie group realized as a group of homeomorphisms of an n-dimensional manifold M of class C^1 (see §3) such that the functions $\bar{x}^i = f^i(x^1, \cdots, x^n; \alpha^1, \cdots, \alpha^r)$ describing locally the effect of the transformation $A = (\alpha)$ on the point $P = (x)$ of M have the derivatives $\partial^2 f^i / \partial x^k \partial \alpha^j$ continuous in (x, α).[3]

3. n-dimensional Riemannian manifolds

Consider an ordinary n-dimensional manifold M; i.e. a connected Hausdorff space each point having a neighborhood which is an n-cell. M is said to be

[1] See, for example, L. P. Eisenhart, *Riemannian Geometry*, (1926), Chapt. 6.

[2] It is usually demanded that the ϕ's be analytic; but Garrett Birkhoff has shown how to introduce analytic parameters in a group of class C^1: *Analytical groups*, Trans. Amer. Math. Soc., vol. 43 (1938), pp. 61–101.

[3] It would be sufficient to require that the α-derivatives exist at the identity and that these derivatives should satisfy a Lipschitz condition in the x's.

of class C^r if to each sufficiently small n-cell neighborhood of each point can be attached a coördinate system such that in the region of overlapping of two neighborhoods the transformation between coördinates of the two neighborhoods is of class C^r and non-singular. M is a *Riemannian manifold of class* C^r ($r \geqq 1$) if it is a manifold of class C^{r+1} and if each coördinate neighborhood is provided with a real, positive definite, symmetric quadratic differential form of class C^r such that in the region of overlapping of two neighborhoods with respective coördinate systems (x) and (y) and quadratic forms $g_{ij}\,dx^i\,dx^j$ and $h_{ij}\,dy^i\,dy^j$ we have

$$g_{ij}\,dx^i\,dx^j = h_{ij}\,dy^i\,dy^j \quad \text{(summation } i, j = 1, \cdots, n\text{)}$$

under the transformation $y^i = y^i(x)$.

Arc length along a curve $x^i = x^i(t)$ is defined in invariant fashion as

$$s = \int \left(g_{ij}\frac{dx^i}{dt}\frac{dx^j}{dt}\right)^{\frac{1}{2}} dt.$$

Here the integral is the Lebesgue integral evaluated along any absolutely continuous representation of a curve admitting such a representation. Such a curve we call an *admissible* curve. By defining the distance ρ between two points as the greatest lower bound of the lengths of admissible arcs joining the points, we can make M into a metric space satisfying the usual axioms. The topology induced by this metric is equivalent to the original manifold-topology. A second kind of length of an arc AB, *metric arc length*, can be defined as the least upper bound of $\sum_0^n \rho(P_i, P_{i+1})$, $(P_0 = A, P_{n+1} = B)$, for all ordered subdivisions $P_1, P_2, \cdots, P_n$ of the arc AB and for all n. If an arc has finite metric arc length, it is said to be *rectifiable*.

It has been shown that on a Riemannian manifold a necessary and sufficient condition that an arc be rectifiable is that the functions defining it be of bounded variation.[4] Thus metric arc length and integral arc length are defined for the same class of arcs, and rectifiable and admissible are equivalent terms. Furthermore it can be proved that the two kinds of arc length are equal.[5]

By a *direction* at P, we shall mean a unit contravariant vector at P. That is, a direction at P associates with each admissible coördinate system (x) containing P: (x_0) a set of components $(u) = (u_1, \cdots, u_n)$ of a contravariant vector, satisfying $g_{ij}(x_0)u^iu^j = 1$. An arc $x^i = x^i(s)$, where s is the arc length measured from P, is said to issue from P in the direction (u) if $x^{i\prime}(0) = u^i$. The *angle* θ between two directions (u) and (v) at P is defined as

$$\cos\theta = g_{ij}(x_0)u^iv^j.$$

The *space of directions* at P is the subspace $g_{ij}(x_0)u^iu^j = 1$ of the n-dimensional number space of the variables (u). This space is homeomorphic to the $(n-1)$-sphere.

---

[4] S. B. Myers, *Arc length in metric and Finsler manifolds*, Annals of Math., vol. 39 (1938), pp. 463–471.

[5] M. Morse, *Functional topology and abstract variational theory*, Memorial des Sc. Math., No. 92 (1938), see p. 67.

Geodesics are extremals in the calculus of variations problem associated with integral arc length. The following are some important local properties of geodesics.[6]

(A) Points on geodesics are functions of class C^r of initial point, initial direction, and length. That is, if g is a geodesic arc $x^i = g^i(s)$, $0 \leqq s \leqq l$, issuing from P: (x_0), the equations of the geodesics issuing from points $(\bar{x})$ near P are given by

$$x^i = X^i(s, \bar{x}, u), \tag{3.1}$$

where

$$X^i(0, \bar{x}, u) \equiv \bar{x}^i, \qquad X^i_s(0, \bar{x}, u) \equiv u^i, \qquad X^i(s, x_0, x_0') \equiv g^i(s).$$

Here X^i and X^i_s are functions of class C^r, s represents arc length, while $(s, \bar{x}, u)$ is a set near (s, x_0, x_0'). In particular, the equations of geodesics through (x_0) are given by

$$x^i = X^i(s, x_0, u). \tag{3.2}$$

(B) The equations (3.1) can also be written as

$$x^i = F^i(\bar{x}, y) \tag{3.3}$$

where $y^i = u^i s$. The functions F^i are of class C^r for $(\bar{x}, y)$ near or at $(x_0, 0)$, and the determinant $|F^i_{y^j}| \neq 0$ and $= \delta^i_j$ at $(\bar{x}, 0)$. Thus the y's can be used as coördinates near P: $(\bar{x})$, and they are called normal coördinates with $(\bar{x})$ as origin. Every point P on the manifold M has a sphere neighborhood in which normal coördinates with P as origin can be used. Such a neighborhood contains no point conjugate to P, is simply covered by the geodesics through P, and these geodesics furnish a proper absolute minimum on M to the length of arcs joining their ends. A neighborhood of this type we shall call an *N-neighborhood* of P.

(C) For every point P there exists a constant $\delta > 0$ such that every sphere neighborhood about P of radius $< \delta$ is an N-neighborhood of all its points. Such a neighborhood will be designated as a *σ-neighborhood* of P.

(D) Let M_r be any regular r-dimensional submanifold of M of class C^r through P: (x_0), with local equations

$$x^i = m^i(\alpha_1, \cdots, \alpha_r), \qquad m^i(0) = x^i_0 \qquad \| m^i_{\alpha_\mu} \| \text{ of rank } r. \tag{3.4}$$

Then the equations of the geodesics perpendicular to M_r at points of M_r near P_0 can be written

$$x^i = M^i(s, \alpha, u), \tag{3.5}$$

[6] See J. H. C. Whitehead, *On the covering of a complete space by the geodesics through a point*, Annals of Math., vol. 36 (1935), pp. 680–685. See also S. S. Cairns, *Normal coördinates for extremals transversal to a manifold*, Amer. Jour. of Math., vol. 60 (1938), pp. 423–435.

where

$$M^i(0, \alpha, u) \equiv m^i(\alpha), \qquad M^i_\alpha(0, \alpha, u) \equiv u^i$$

for (α) near (0) and (u) satisfying

$$g_{ij}[m(\alpha)]m^i_{\alpha_\rho}u^j = 0. \tag{3.6}$$

The functions M^i are of class C^r.

(E) The equations of geodesics perpendicular to M_r can also be written as

$$x^i = N^i(\alpha, z) \qquad (z) = (z_{r+1}, \cdots, z_n) \tag{3.7}$$

where $N^i(\alpha, 0) \equiv m^i(\alpha)$. The functions N^i are of class C^r for (α, z) near (0, 0) and the jacobian of (3.7) does not vanish. For (α) constant, the z's are normal coördinates in the geodesic $(n - r)$-dimensional surface perpendicular to M_r at (α).

To complete our summary of properties of Riemannian manifolds, we state a useful definition and a few lemmas.

Let P_k be a set of points converging to P, and let d_k denote a direction at P_k. Then we shall say that the directions d_k *converge to a limit direction* d at P if no matter what coördinate system (x) about P we use, the components $(u)_k$ of d_k in that system converge to the components (u) of d. To test for such convergence it is sufficient to use just one coördinate system; for the property $(u)_k \to (u)$ is clearly preserved under contravariant transformations.

LEMMA 1. *Every sequence of directions at a convergent sequence of points has a convergent subsequence.*

For if $\sum_{ij} g_{ij}(x_k)u^i_k u^j_k = 1$, for all k as $k \to \infty$ and $(x_k) \to (x_0)$, then standard properties of positive definite quadratic forms show that (u_k) remains bounded.

LEMMA 2. *If a sequence of functions $x^i = x^i_k(s)$ representing geodesics arcs converges as $k \to \infty$ to a function $x^i = x^i_0(s)$ representing a geodesic arc, then $x^{i\prime}_k(s) \to x^{i\prime}_0(s)$.*

This follows from Lemma 1 and (A).

LEMMA 3. *For a compact set N covered by a coördinate system (x), corresponding to an arbitrary $\epsilon > 0$ there exists a $\delta > 0$ independent of Q in N such that if two directions at Q form an angle $< \delta$, their respective components (u) and (v) satisfy $|v^i - u^i| < \epsilon$.*

For if (u) and (v) form an angle γ, we have

$$g_{ij}(x)(v^i - u^i)(v^j - u^j) = 2(1 - \cos\gamma) = 4\sin^2\frac{\gamma}{2}.$$

Properties of positive definite quadratic forms enable one to see that $|v^i - u^i|$ can be made arbitrarily small by suitable restriction of γ.

Any property of Riemannian manifolds stated and used later without proof is provable from properties (A)–(E) and lemmas (1)–(3) of this section.

4. Isometries

An *isometry* of a metric space is defined as a distance preserving homeomorphism of the space into itself. On the other hand, one might naturally

define an isometry of a Riemannian manifold M as a homeomorphism of M into itself preserving integral arc length.

THEOREM 1. *The distance preserving homeomorphisms of a Riemannian manifold M are identical with the homeomorphisms preserving integral arc-length.*

For if a homeomorphism of M into itself preserves distance, it preserves metric arc length, hence integral arc length. Also, if a homeomorphism preserves integral arc length, it preserves distance, since the latter is defined as the greatest lower bound of integral arc length.

We note from the minimizing properties of short geodesic arcs that an isometry carries geodesics into geodesics.

THEOREM 2. *An isometry of a Riemannian manifold of class C^1 is a transformation of class C^1.*

Necessary and sufficient conditions that a rectifiable arc C: $x^i = x^i(s)$ have a derivative $x^{i'}$ with respect to s at a point P are the following:

(1) The direction (p) of the geodesic PQ at P converges to a direction (p_0) as $Q \to P$ on C.

(2) Lim $(\Delta r/\Delta s)$ exists, where $\Delta r = \rho(P, Q)$ and $\Delta s =$ length of PQ on C.

The sufficiency of these conditions follows from the fact that using normal coördinates (y) about P, and denoting by (Δy) the normal coördinates of Q, we have

$$\frac{\Delta y^i}{\Delta s} = \frac{\Delta y^i}{\Delta r} \cdot \frac{\Delta r}{\Delta s} = p^i \frac{\Delta r}{\Delta s}. \tag{4.1}$$

The necessity of conditions is proved as follows. Let $h_{ij}(y)$ be the coefficients of the fundamental quadratic form in the coördinates (y). Then

$$\Delta r = \Delta r(h_{ij}(0)p^i p^j)^{\frac{1}{2}} = (h_{ij}(0)p^i \Delta r p^j \Delta r)^{\frac{1}{2}} = (h_{ij}(0)\Delta y^i \Delta y^j)^{\frac{1}{2}}.$$

Hence

$$\lim \frac{\Delta r}{\Delta s} = \left(h_{ij}(0)\, \frac{dy^i}{ds}\, \frac{dy^j}{ds}\right)^{\frac{1}{2}}.$$

From (4.1) it follows that lim p^i also exists.

Now the first of the two conditions is invariant under an isometry because of the continuity in points and directions of the transformations of geodesics effected by an isometry. The second condition is obviously invariant. Hence if C has a direction at P, its image $\bar{C}$ has a direction at $\bar{P}$, so that we may say that an isometry carries a direction into a direction. Furthermore, the new direction varies continuously with the old direction and the point at which it is taken.

Under these circumstances, an isometry must be of class C^1. For consider the curve

$$x^1 = a(s, c), \quad (c) = (c^2, \cdots, c^n); \quad x^k = c^k \quad (k = 2, \cdots, n), \tag{4.2}$$

where $a(s, c)$ is obtained by solving

$$s = \int_{c^1}^{x^1} (g_{11}(x^1, c^2, \cdots, c^n))^{\frac{1}{2}}\, dx^1 \tag{4.3}$$

for x^1. The function $a(s, c)$ is of class C^1. Hence after performing the isometry $y^i = y^i(x)$, the curve $y^i = y^i[a(s, c), c]$ has a derivative D with respect to s. But $\partial a/\partial s$ exists, hence $\partial y^i/\partial x^1$ exists, and

$$D = \frac{\partial y^i}{\partial x^1}\,[a(s, c), c]\,\frac{\partial a}{\partial s}\,(s, c).$$

Furthermore D is continuous in (s, c). Also $\partial a/\partial s$ is continuous in (s, c) and $\neq 0$. Hence $\partial y^i/\partial x^1$ is continuous in (s, c). But (4.3) is continuous in x^1 and c; hence $\partial y^i/\partial x^1$ is continuous in (x^1, c). Similarly $\partial y^i/\partial x^k$ $(k = 2, \cdots, n)$ exists and is continuous.

An immediate consequence of this theorem is that an isometry transforms vectors and tensors according to the usual laws, transforms the space of directions at a point linearly, and preserves angles.

THEOREM 3. *If an isometry T leaves fixed $n + 1$ points so close together that n of them lie in an N-neighborhood of the other, and if the points are linearly independent (i.e. not in the same $(n - 1)$-dimensional geodesic hypersurface), then T is the identity.*

Suppose T leaves fixed the $n + 1$ linearly independent points $P_0, P_1, \cdots, P_n$, of which $P_1, \cdots, P_n$ lie in the N-neighborhood σ of P_0. Then the short geodesics $P_0P_1, P_0P_2, \cdots, P_0P_n$ are fixed under T, so are their initial directions, and hence the linear homogeneous transformation of the space of directions at P_0 induced by T is the identity. Hence all geodesic arcs issuing from P_0 are fixed; since length is preserved, they are pointwise fixed. Thus any N-neighborhood of P_0 is pointwise fixed. If Q is a point in an N-neighborhood of P_0, any N-neighborhood of Q is similarly pointwise fixed. But P_0 can be joined to any point P in M by a broken geodesic arc with a finite number of pieces, each corner being in an N-neighborhood of the preceding corner. Thus P is fixed.

COROLLARY. *There is at most one isometry which carries $n + 1$ points P_i of the kind described in the theorem into $n + 1$ points Q_i.*

For if we had two such isometries T and T', then by the theorem, $T^{-1}T'$ would leave fixed P_i, and so would be the identity.

5. The group of isometries and its topology

The isometries of M form a group, which we shall call G. A sequence of isometries T_k will be said to converge to an isometry T if for every P in M, $T_k(P) \to T(P)$. With this notion of convergence, the group can be shown

to be *continuous*. It can also be shown[7] that if $T_k(P)$ converges to a point $\overline{P}$, there exists a subsequence T_k' of T_k converging to an isometry T such that $T(P) = \overline{P}$. If $T_k(P)$ converges to P, T leaves P fixed; the sequence $T_k'T^{-1}$ takes P into the same set of points as T_k', and converges to the identity. It follows that if M is compact, G is compact (i.e. every sequence of isometries in G has a subsequence converging to an isometry in G). In these properties, G could be replaced by a closed subgroup H of G.

Without difficulty, we can now topologize, in fact metrize, G so that we obtain the above notion of convergence. Consider any set of $n + 1$ points of the kind used in Theorem 3. Then the distance $d(T, \overline{T})$ between two isometries T and $\overline{T}$ will be defined as the maximum of the distance $\rho[T(P), \overline{T}(P)]$ as P ranges over the given set of $n + 1$ points. This distance can be shown to satisfy the usual metric axioms. Obviously the previous notion of convergence of isometries T_i to T implies $d(T_i, T) \to 0$; we must show that $d(T_i, T) \to 0$ implies $T_i(P) \to T(P)$ for all P on M. This will not only show that convergence in this new topology is equivalent to the previous notion of convergence, but also that the new topology is independent of the particular set of $n + 1$ points used.

If it were not true that from $d(T_i, T) \to 0$ follows $T_i(P) \to T(P)$ for all P, there would exist a point Q, an $\epsilon > 0$, and a subsequence T_i' of T_i such that

$$\rho[T_i'(Q), T(Q)] > \epsilon \qquad \text{for all } i. \tag{5.1}$$

But we know that $T_i'(P) \to T(P)$ for all P on the given set of $n + 1$ points; hence T_i' must contain a subsequence T_i'' such that there exists an isometry T' with $T_i''(P) \to T'(P)$ for all P. Now $T'(P) = T(P)$ for P on the given set of $n + 1$ points; hence by the corollary to Theorem 3, $T' = T$. This contradicts (5.1), and the proof is complete.

We could equally well have defined $d(T, \overline{T})$ as the maximum of $\rho[T(P), \overline{T}(P)]$ for all P on a compact subset of M containing an open set. If M is compact, an obvious choice of the subset is M itself.

LEMMA 4. *If a sequence of isometries T_k converges to an isometry T, as a sequence of transformations of the space of directions at any point P_0 it converges uniformly.*

Without loss of generality, assume T = identity. Let $T_k(P_0) = P_k$. Then T_k transforms each geodesic arc issuing from P_0 in a direction $(u)_0$ into a sequence of geodesic arcs issuing from P_k and satisfying the hypotheses of Lemma 2, hence $(u)_k = T_k(u)_0 \to (u)_0$. Also, T_k transforms the compact space of directions at P_0 in a one-to-one continuous manner and this continuity of T_k in direction is uniform in k. That is, for an arbitrary $\epsilon > 0$ there exists a $d > 0$ independent of k such that, if $| v_0^i - u_0^i | < d$, then $| v_k^i - u_k^i | < \epsilon$. For we simply choose d so small that, if $| v_0^i - u_0^i | < d$, then the angle between

[7] Van Dantzig and van der Waerden, *Über metrisch homogene Räume*, Abh. Math. Sem. Hamburg, vol. 6 (1928), pp. 374-376.

$(v)_0$ and $(u)_0$ is less than the δ of Lemma 3; the angle between $(v)_k$ and $(u)_k$ will also be less than δ, and Lemma 3 tells us that $|v_k^i - u_k^i| < \epsilon$ independent of k (for k large enough).

Now let ϵ be arbitrary again. By what we have just shown, we can choose a finite set of directions at P_0, $\epsilon/3$-dense in the direction space at P_0, whose transforms at P_k are also $\epsilon/3$-dense in the direction space at P_k. Now let N be taken so large that for $k > N$ each of our finite set of directions at P_k, say $(v)_k$ is within $\epsilon/3$ of its limit direction at P_0, i.e. $|v_k^i - v_0^i| < \epsilon/3$ for $k > N$. But for an arbitrary direction $(u)_k$ at P_k, we have (for one of the $(v)_k$'s) $|u_k^i - v_k^i| < \epsilon/3$ and $|u_0^i - v_0^i| < \epsilon/3$. Thus $|u_k^i - u_0^i| < \epsilon$ for $k > N$. This proves the lemma.

6. The set of equivalent points

Two points are said to be *equivalent* under a closed subgroup H of G if there exists an isometry of H taking one point into the other. According to §5, the set γ of points equivalent under H to a point P_0 of M is closed. We shall find properties of the group H by a study of the set of points γ.

The point P_0 is said to be a *limit point of equivalent points P_k from a direction* (u) if $P_k \to P_0$ and the directions at P_0 of the short geodesics P_0P_k converge to the direction (u).

THEOREM 4. *If P_0: (x_0) is a limit point of equivalent points P_k from a direction (u_0), it is also a limit point of equivalent points from the opposite direction $(-u_0)$.*

Let us pass immediately to a sequence T_k such that the isometries T_k taking P_0 into P_k converge to the identity. We consider three sets of directions: the direction (u_k) of the geodesic P_0P_k at P_0, the direction (v_k) of the geodesic P_0P_k at P_k, and the direction (w_k) at P_0 which is transformed into (v_k) by T_k. We note that $(u_k) \to (u_0)$. The equations of the geodesics through P_0 are given by (3.2). If we differentiate with respect to s, we obtain

$$x^{i\prime} = X_s^i(s, x_0, u). \tag{6.1}$$

Now $v_k^i = X_s^i(s_k, x_0, u_k)$ and $u_0^i = X_s^i(0, x_0, u_0)$ where s_k is the distance P_0P_k. As $k \to \infty$, $s_k \to 0$, and $(u_k) \to (u_0)$. The continuity of the functions (6.1) tells us that $(v_k) \to (u_0)$.

But by Lemma 4, if k is large enough, $|v_k^i - w_k^i| < \epsilon$ for a given $\epsilon > 0$. Hence (w_k) also converges to (u_0). Now the geodesic h_k issuing from P_0 in the direction (w_k) is transformed by T_k into the extension of the geodesic P_0P_k. Hence there is a point on h_k equivalent to P_0 at a distance of $\rho(P_0, P_k)$ before it. Thus P_0 is a limit of equivalent points from the direction $(-u_0)$.

THEOREM 5. *The set of point of M equivalent to P_0 under a closed subgroup H of G is locally euclidean or discrete.*

First, if the set γ_n of points near P_0 equivalent to P_0 is an n-cell, the theorem is true. The total set of points equivalent to P_0 is, in this case, the whole of M; for the set is homogeneous, hence open in M, and at the same time closed in M.

In general, arbitrarily close to P_0 is a point not equivalent to P_0 and hence

not a limit point of points equivalent to P_0. Let Q be such a point in a small σ-neighborhood σ of P_0. (Henceforth everything will take place inside σ). Let r be the distance from Q to a nearest point A equivalent to P_0. The set of points P such that $\rho(Q, P) < r$ contains no point equivalent to P_0. The locus Σ of points P such that $\rho(Q, P) = r$ will be entirely within σ if Q was taken close enough to P_0. The set Σ is homeomorphic to an $(n-1)$-sphere, and is given by equations

$$x^i = X^i(r, x_Q, u) = x^i(u) \tag{6.2}$$

where

$$g_{ij}(x_Q)u^iu^j = 1. \tag{6.3}$$

These are obtained from (3.1) by substituting $(\bar{x}) = (x_Q)$, $s = r$. Locally we can put (6.3) into parametric form $u^i = u^i(\alpha_1, \cdots, \alpha_{n-1})$, obtaining a local representation of class C^1 for Σ: $x^i = x^i[u(\alpha)] = f^i(\alpha)$. This representation is regular; i.e., $\| f^i_{\alpha_\mu} \|$ is of rank $n-1$; otherwise we would have $\det(\| X^i_s \| \cdot \| X^i_{u^j} u^i_{\alpha_\mu} \|) = 0$ contradicting the fact that Q has no conjugate point in σ. Hence at every one of its points, in particular at A, Σ is a regular manifold of class C^1 and has a tangent $(n-1)$-dimensional space τ of directions.

Now it can be seen that A cannot be a limit point of equivalent points from any direction but those of τ. For suppose the contrary, then by Theorem 4, A is a limit point of equivalent points from two opposite directions. The geodesic ray g issuing from A in one of these two directions, say (u), goes into the interior of Σ; it lies within this interior for a certain length d. Geodesics from A in directions in a small neighborhood of (u) have the same property with the same number d. Since A is a limit point of equivalent points from the direction (u), there exists a sequence of equivalent points A_k converging to A such that the directions at A of the geodesics AA_k converge to (u); this sequence must be interior to Σ, contradicting the fact that A was the nearest point to Q equivalent to P_0.

Thus the point A can be a limit point of equivalent points only from directions in an $(n-1)$-dimensional space of directions, and the same statement must apply to P_0. Let τ_{n-1} be this space of directions at P_0, and let H_{n-1} be a regular $(n-1)$-dimensional submanifold of M of class C^2 and tangent to τ_{n-1} at P_0. We will set up a homeomorphism between the set γ_n of points equivalent to P_0 near it in M and a subset of the points of H_{n-1} near P_0.

The geodesics issuing from points of H_{n-1} near P_0 perpendicular to H_{n-1} simply cover a neighborhood of P_0 in M according to (E). To every point P equivalent to P_0 in a sufficiently small neighborhood of P_0 we make correspond the foot of the unique perpendicular geodesic from P to H_{n-1}. We will show that this correspondence C is $1-1$ in a small neighborhood of P_0.

If two points P and Q equivalent to P_0 correspond under C to the same point B of H_{n-1}, P and Q both lie on the geodesic through B perpendicular to H_{n-1}. If this situation occurs in an arbitrarily small neighborhood of P_0,

there exists a sequence of points P_k equivalent to P_0 converging to P_0, each of which lies on a geodesic P_kQ_k, perpendicular to H_{n-1}, containing another point Q_k equivalent to P_0, while $\rho(P_k, Q_k) \to 0$. Choosing a sequence of isometries T_k converging to the identity such that T_k takes P_0 into P_k, let h_k be the geodesic that T_k takes into P_kQ_k. Then by Lemma 4, if k is large enough the difference between the direction of h_k at P_0 and the direction of P_kQ_k at H_{n-1} will be less than ϵ (arbitrary). Also, for k large enough, the difference between the direction of P_kQ_k at H_{n-1} and its direction at P_k is $<\epsilon$. This follows from the fact that the functions $\partial M^i/\partial s$ of (3.5) are continuous in s uniformly in (α). The directions of a subsequence $P_k'Q_k'$ of the geodesics P_kQ_k at H_{n-1} converge to a direction (u) perpendicular to H_{n-1} at P_0. Hence the directions of $P_k'Q_k'$ at P_k', and also the directions of h_k' at P_0, converge to (u). Thus P_0 is a limit point of equivalent points $R_k = T_k^{-1}(Q_k')$ from a direction perpendicular to H_{n-1} at P_0, which contradicts the definition of H_{n-1}.

Hence there exists a compact neighborhood σ_n of P_0 in M in which the set $\bar{\gamma}_n$ of points equivalent to P_0 is in a $1-1$ correspondence C with a subset $\bar{\gamma}_{n-1}$ of an $(n-1)$-cell in H_{n-1} containing P_0. As a transformation from $\bar{\gamma}_n$ to $\bar{\gamma}_{n-1}$, C is continuous. Since $\bar{\gamma}_n$ is compact, C is bicontinuous, hence a homeomorphism.

We will say that the set γ_n has the property A_{n-p} $(1 \leqq p \leqq n)$ if

(1) P_0 is a limit point of equivalent points from at most an $(n-p)$-dimensional space of directions τ_{n-p}; and

(2) γ_n is homeomorphic to a subset γ_{n-p} of an $(n-p)$-dimensional regular submanifold H_{n-p} of M of class C^2 tangent to τ_{n-p} at P_0, the homeomorphism being obtained by making a point P of γ_n correspond to the foot of the perpendicular geodesic from P to H_{n-p}.

We have already proved that either γ_n is an n-cell, or else it has property A_{n-1}. We now prove generally that if γ_n has property A_{n-p}, either it is an $(n-p)$-cell, or else it has property A_{n-p-1}.

Assume γ_n has property A_{n-p}, but is not an $(n-p)$-cell. There will be a sequence of points P_k in H_{n-p} converging to P_0 no one of them in γ_{n-p}. The set γ_{n-p} is closed (since γ_n is closed). Hence around each P_k can be found a sphere neighborhood in H_{n-p} containing no points of γ_{n-p}. Let r_k be the upper bound of the radii of such sphere neighborhoods of the point P_k; then the $(n-p-1)$-sphere R_k of radius r_k about P_k in H_{n-p} has a point W_k of γ_{n-p} on it, but no points of γ_{n-p} interior to it. Erect all geodesics in M in directions perpendicular to H_{n-p} at the points of this sphere and its interior σ_k. There will be p linearly independent such directions at each point. By (E), if P_k is close enough to P_0, these geodesics simply cover an n-dimensional cell. The geodesics perpendicular to H_{n-p} at the points of R_k form an $(n-1)$-dimensional cylindrical boundary S_k of the n-cell (topologically the product of an $(n-p-1)$-sphere by a p-plane). Then S_k contains a point Z_k equivalent to P_0, but in the interior of the n-cell bounded by S_k are no points equivalent to P_0. Using (D) and (E), it can be shown that S_k is a regular manifold of

class C^1. Following reasoning used previously, Z_k can be a limit point of equivalent points at most from directions in the $(n-1)$-plane of directions λ_k tangent to S_k at Z_k.

Now as $k \to \infty$, a subsequence λ_k' of λ_k will converge to an $(n-1)$-dimensional space of directions λ_0 at P_0 containing p linearly independent directions perpendicular to H_{n-p} at P_0. Let T_k be a sequence of isometries converging to the identity and taking P_0 into Z_k', and let ξ_k be the $(n-1)$-dimensional space of directions at P_0 which T_k takes into λ_k'. The point P_0 can be a limit point of equivalent points at most from directions in ξ_k. By Lemma 4, since λ_k' converges to λ_0, so does ξ_k. Choose a value of k so large that ξ_k contains p linearly independent directions each almost perpendicular to H_{n-p} at P_0. Then the intersection of ξ_k and τ_{n-p} (the plane of directions tangent to H_{n-p} at P_0) is at most $(n-p-1)$-dimensional. Thus we have proved that P_0 is a limit point of equivalent points from at most an $(n-p-1)$-dimensional space of directions τ_{n-p-1}, the first part of property A_{n-p-1}.

Now let H_{n-p-1} be a regular $(n-p-1)$-dimensional manifold of class C^2 tangent to τ_{n-p-1} at P_0. Set up the correspondence C between points of γ_n and the feet of the geodesics drawn from the points of γ_n perpendicular to H_{n-p-1}. If C is not $1-1$ in some small neighborhood of P_0, there would be a sequence of pairs of points P_k and Q_k of γ_n converging to P_0 such that the geodesic drawn from P_k perpendicular to H_{n-p-1} and the geodesic drawn from Q_k perpendicular to H_{n-p-1} would intersect H_{n-p-1} at the same point D_k. Thus P_k and Q_k would lie on the geodesic $(p+1)$-surface G_k perpendicular to H_{n-p-1} at D_k. Draw the geodesic P_kQ_k.

Now we shall show that a subsequence of the directions (u_k) of P_kQ_k at P_k converges to a direction at P_0 tangent to G_0, the geodesic surface perpendicular to H_{n-p-1} at P_0. Let us use the coördinates (α, z) of (E) (§3) based on H_{n-p-1} at P_0. The equations of G_k are $\alpha_\mu = \text{constant}$. Writing the equations of the geodesics P_kQ_k as

$$\alpha_\mu = \alpha_\mu(s), \qquad z_r = z_r(s),$$

by the law of the mean, at some point on P_kQ_k between P_k and Q_k we must have $\alpha_\mu' = 0$. As $k \to \infty$, P_k and Q_k approach P_0, $\rho(P_k, Q_k) \to 0$, and (u_k) has at least one limiting direction (u_0) at P_0. By the continuity of the functions (6.1), the first $n-p-1$ components of (u_0) must be zero (in the coördinate system (α, z)). Hence (u_0) is tangent to G_0, and so a subsequence of (u_k) converges to a direction perpendicular to H_{n-p-1}.

Following the usual procedure, we now let T_k be a sequence of isometries converging to the identity and taking P_0 into P_k, and let (v_k) denote the direction which T_k takes into (u_k). Then by Lemma 4, since (u_k) converges to (u_0), so does (v_k). But this means that P_0 is a limit point of equivalent points $B_k = T_k^{-1}(Q_k)$ from the direction (u_0) perpendicular to H_{n-p-1} at P_0, contradicting the definition of H_{n-p-1}.

Thus the correspondence C between points of γ_n and points of a subset

γ_{n-p-1} of H_{n-p-1} is $1-1$. It is easily shown to be continuous, and so is a homeomorphism.

This completes the proof that if γ_n has property A_{n-p}, then either it is an $(n-p)$-cell, or else it has property A_{n-p-1}. But this brings to a conclusion the proof of Theorem 5. For it is now clear that either γ_n is a q-cell for some q between 0 and n, or else it has property A_0. The latter means that γ_n consists entirely of P_0. Since the total set of points equivalent to P_0 is homogeneous it is everywhere locally euclidean or it is discrete.

In order to prove that γ_n is a manifold of class C^1 we shall need the following lemma.

LEMMA 5. *Let $x^i = x^i(t)$ be an arc C issuing from a point $P_0 : (x_0)$. Let Q be a nearby point on C, with coördinates $(x_0 + \Delta x)$, and let Δr be the distance from P_0 to Q. Then if the direction (p) of the geodesic P_0Q converges as $Q \to P_0$ to a unique limiting direction (p_0) at P_0, the ratio $\Delta x^i/\Delta r$ approaches p_0^i.*

Transform from the coördinates (x) to normal coördinates (y) with P_0 as origin. Call the new coördinates of Q (Δy). The components of directions issuing from P_0 are the same in the coördinate system (y) as in the system (x). Now $\Delta y^i = p^i \Delta r$. Therefore $\lim (\Delta y^i/\Delta r) = \lim p^i = p_0^i$. Transform back to the coördinates (x), and use the fact that the jacobian of the transformation at P_0 is the unit matrix.

THEOREM 6. *The set of points of M equivalent to P under a closed subgroup H of G is a regular submanifold of M of class C^1.*

Let P_0 be an arbitrary point of the manifold S of equivalent points. Suppose S is r-dimensional. We know that P_0 is a limit point of S only from directions in a certain r-plane τ_r of directions at P_0. Let M_r be a regular r-dimensional submanifold of M of class C^2 tangent to τ_r at P_0. Transform from given coördinates (x) to the coördinates (α, z) of (E) (see §3), based on M_r about P_0. The equations of M_r are $z^{r+1} = \cdots = z^n = 0$, while $(\alpha) =$ constant represents a geodesic $(n-r)$-dimensional manifold normal to M_r. We know from the proof of Theorem 5 that there is a homeomorphism between points of S and M_r neighboring P_0 in which a point of S corresponds to the foot of the perpendicular geodesic from it to M_r. Corresponding points in S and M_r now have the same (α) coördinates. We can use (α) as parameters in S, and write the equations of S as

$$\alpha^\mu = \alpha^\mu \quad (\mu = 1, \cdots, r), \qquad z^\tau = z^\tau(\alpha) \quad (\tau = r+1, \cdots, n).$$

Let $\bar{P}$: $(\bar{\alpha}, \bar{z})$ be any point of S near P_0. Consider the curve C on S through $\bar{P}$ defined by

$$\alpha^1 = t, \qquad \alpha^\rho = \bar{\alpha}^\rho \ (\rho = 2, \cdots, r), \qquad z^\tau = z^\tau\,(t, \bar{\alpha}^2, \cdots, \bar{\alpha}^r), \qquad (t \text{ near } \bar{\alpha}^1).$$

Let Q be a nearby point on C. Denote the geodesic $\bar{P}Q$ by g, its direction at $\bar{P}$ by (h). Let $(\bar{h})$ be any limiting direction of (h) as $Q \to \bar{P}$; we shall prove that $(\bar{h})$ is unique.

By the law of the mean, at some point on g between $\bar{P}$ and Q the ρ^{th} com-

ponent ($\rho = 2, \cdots, r$) of the direction of g is zero. Letting $Q \to \bar{P}$, and using the continuity of (6.1), it follows that $\bar{h}^\rho = 0$ ($\rho = 2, \cdots, r$). Also we know that $(\bar{h})$ must lie in a certain r-plane $\bar{\tau}_r$ of directions at $\bar{P}$. This r-plane can be made arbitrarily close to the corresponding r-plane τ_r at P_0 by taking $\bar{P}$ close enough to P_0, using Lemma 4 in the usual way. The conditions that a direction (d) at P_0 lie in τ_r are $d^{r+1} = \cdots = d^n = 0$. Hence the $n - r$ linear homogeneous conditions that $(\bar{h})$ lie in $\bar{\tau}_r$ must also be linearly independent on $(\bar{h}^{r+1}, \cdots, \bar{h}^n)$.

Thus $\bar{h}^2, \cdots, \bar{h}^n$ are determined in terms of $\bar{h}^1$, which must be $\neq 0$. The fact that $(\bar{h})$ is a unit vector furnishes the proof of its uniqueness.

We now apply Lemma 5 which gives $\lim (\Delta z^r/\Delta r) = \bar{h}^1$ as $Q \to \bar{P}$. This together with $\lim (\Delta\alpha^1/\Delta r) = \bar{h}^1 \neq 0$ proves the existence of $\partial z^r/\partial\alpha^1$ at $(\bar{\alpha})$. Similarly $\partial z^r/\partial\alpha^\rho$ exists.

The continuity of these derivatives follows from the facts already mentioned concerning the continuous variation of $\bar{\tau}_r$ with $(\bar{\alpha})$.

Thus the $z^r(\alpha)$ are of class C^1 near P_0. We can transform back to the original coördinates (x) and obtain a local representation of S: $x^i = x^i(\alpha)$ of class C^1, with functional matrix of rank r. This proves Theorem 6.

7. Closed groups of isometries

Let M^{n+1} be the product manifold of M taken with itself $n + 1$ times. A point of M^{n+1} is an ordered set of $n + 1$ points of M. Coördinate neighborhoods in M^{n+1} are the products of $n + 1$ coördinate neighborhoods in M and the fundamental form in such a neighborhood is the sum of the quadratic forms of the corresponding neighborhoods in M. Thus M^{n+1} is a Riemannian manifold of the same class as M. Furthermore, if T is an isometry of M, the transformation

$$T(P_0, P_1, \cdots, P_n) = (T(P_0), T(P_1), \cdots, T(P_n)) \tag{7.1}$$

is an isometry of M^{n+1}. Thus if H is a group of isometries of M we can also consider it as a group of isometries of M^{n+1}. If H is a closed group of isometries of M, it is a closed group of isometries of M^{n+1}. This follows from the fact that any limit of isometries of type (7.1) is again of this type.

Let us choose $P_0, P_1, \cdots, P_n$ in a σ-neighborhood of P_0 in M so that the only isometry of M leaving them fixed is the identity. Then if H is a set of isometries of M, and γ is the set of points of M^{n+1} into which $\bar{P} = (P_0, P_1, \cdots, P_n)$ is mapped by H, the transformation sending $T \in H$ into $T(\bar{P}) \in \gamma$ is a $1 - 1$ mapping of H onto all of γ. This mapping is bicontinuous. For clearly $T_k \to T$ implies $T_k(\bar{P}) \to T(\bar{P})$, since $T_k(P_i) \to T(P_i)$. Conversely if $T_k(\bar{P}) \to T(\bar{P})$, from any subsequence of $\{T_k\}$ we can choose a new subsequence $\{T_k'\}$ converging to a limit T_0; and as $T_0(\bar{P}) = T(\bar{P})$ we must have $T_0 = T$. This implies that $T_k \to T$. Thus we have proved

LEMMA 6. *The transformation sending $T \in H$ into $T(\bar{P})$ in the product manifold M^{n+1}, where $\bar{P} = (P_0, P_1, \cdots, P_n)$ is a set of points of M as in Theorem 3.*

is a $1-1$ *bicontinuous transformation of the set* H *of isometries of* M *onto the set of points into which* $\overline{P}$ *is mapped by* H.

As an immediate consequence we have

THEOREM 7. *If* H *is a closed group of isometries of* H, *it is in* $1-1$ *bicontinuous correspondence with a manifold of class* C^1 *imbedded in* M^{n+1}. *Hence* H *is locally euclidean.*

8. Lie groups of isometries

Our object in this section is to show that the parameters introduced in H in §7 make of H a Lie group of transformations of M. We shall need the following results.

Suppose, under the isometry T_0, the point P is transformed into $T_0(P) = Q$. Let U be a σ-neighborhood of Q with coördinates (y). Choose a σ-neighborhood V of P with coördinates (x) such that $T_0(\overline{V}) \subset U$. Then, if T is in a neighborhood W of T_0, this transformation is given by functions

$$y^i = f^i(x^1, \cdots, x^n; T) = f^i(x; T) \tag{8.1}$$

which are continuous in $(x; T)$ by definition of the topology in the group of isometries. We now consider the differentiability properties of these functions.

If M is of class C^r (analytic) the functions (3.1) $x^i = X^i(r, \bar{x}, u)$ defined in V are of class C^r (analytic). Locally the equations $g_{ij}(\bar{x})u^iu^j = 1$ (g_{ij} is the fundamental form in V) can be put in parametric form

$$u^i = u^i(\alpha_1, \cdots, \alpha_{n-1}) \tag{8.2}$$

obtaining a representation of class C^r (analytic) of (3.1)

$$x^i = X^i(r, \bar{x}, u(\alpha)). \tag{8.3}$$

As there is no pair of conjugate points in V, $\det(\| X^i_r \| \cdot \| X^i_{u^j} u^j_{\alpha_\rho} \|) \neq 0$. Hence we can solve (8.3) obtaining functions $r(\bar{x}, x)$ and $\alpha_\rho(\bar{x}, x)$ of class C^r (analytic) for $(\bar{x})$ and (x) in V, $(\bar{x}) \neq (x)$. Substituting these latter in (8.2) we obtain

$$u^i(\bar{x}, x) = u^i(\alpha(\bar{x}, x)) \tag{8.4}$$

of class C^r (analytic).

Choose a point $P_0 \neq P$ in V and let $g_1, \cdots, g_n$ be a set mutually orthogonal geodesics issuing from P_0. Choose a point P_k on g_k so that the part of g_k joining P_0 to P_k is in V. The cosine of the angle between the geodesics P_0P and P_0P_k is $g_{ij}(x_0)u^i(x_0, x)u^j(x_0, x_k)$ where (x) are the coördinates of P and (x_k) are coördinates of P_k. These cosines we denote by $\pi_k(x)$; they are of class C^r (analytic).

In the neighborhood U of Q let the equations (3.1) of the geodesics be written $y^i = Y^i(r, \bar{y}, v)$ again of class C^r (analytic). Corresponding to (8.4) the direction of the geodesic from $(\bar{y})$ to (y) is given by functions $v^i(\bar{y}, y)$. Set

$$v^i_k(T) = v^j[f(x_0; T); f(x_k; T)]. \tag{8.5}$$

Denote by $\bar{g}_{ij}(y)$ the fundamental form in U.

Since isometries preserve angles (§4), the direction (v) of the geodesic from $f(x_0\,;\,T)$ to $f(x;\,T)$ must satisfy

$$\bar{g}_{ij}[f(x_0\,;\,T)]v^i v_k^j(T) = \pi_k(x). \tag{8.6}$$

Since $g_1, \cdots, g_n$ are mutually orthogonal, $Tg_1, \cdots, Tg_n$ are also. Hence $\det\{\bar{g}_{ij}[f(x_0\,;\,T)]v_k^j(T)\} \neq 0$. We can therefore solve (8.6) for the v^i obtaining them as linear functions of the $\pi_k(x)$:

$$v^i = a^{ik}(T)\pi_k(x). \tag{8.7}$$

It is to be observed that $a^{ik}(T)$ is of class C^r (analytic) in the functions $f(x_0\,;\,T)$, $\cdots, f(x_n\,;\,T)$.

Since isometries preserve distance $r(x_0\,;\,x) = \bar{r}[f(x_0\,;\,T);\,f(x;\,T)]$ where $\bar{r}(\bar{y},\,y)$ is the distance from $(\bar{y})$ to (y). Since the point $T(P)$ must bear the same relation to $Tg_1, \cdots, Tg_n$ as P bears to $g_1, \cdots, g_n$ we obtain the following representation of the functions $f^i(x;\,T)$:

$$f^i(x;\,T) = Y^i[r(x_0\,;\,x),\,f(x_0\,;\,T),\,a^k(T)\pi_k(x)]. \tag{8.8}$$

Since Y^i, r, π_k are of class C^r (analytic) in (x) we obtain

LEMMA 7. *The functions $f^i(x;\,T)$ are of class C^r (analytic) in (x) and their derivatives are continuous in $(x;\,T)$.*

As an immediate consequence we have

THEOREM 8. *An isometry of a Riemannian manifold of class C^r (analytic) is of class C^r (analytic).*[8]

We shall draw some further consequences of this representation. Suppose the neighborhood W is given coördinates $(\beta^1, \cdots, \beta^r) = (\beta)$ so that $f^i(x;\,T)$ can be written $f^i(x;\,\beta)$. Suppose furthermore that the derivatives $\partial f^i/\partial\beta^k$ $(i = 1, \cdots, n;\ k = 1, \cdots, r)$ exist at the points $P_0, \cdots, P_n$ for all (β) in W. We can then differentiate (8.8) obtaining

$$\frac{\partial f^i(x;\,\beta)}{\partial\beta^k} = \frac{\partial Y^i}{\partial\bar{y}^j}\,\frac{\partial f^j(x_0\,;\,\beta)}{\partial\beta^k} + \frac{\partial Y^i}{\partial v^j}\,\pi_l(x)\,\frac{\partial a^{jl}(\beta)}{\partial\beta^k}. \tag{8.9}$$

Since a^{jl} is of class C^r in $f(x_0\,;\,\beta), \cdots, f(x_n\,;\,\beta)$, $\partial a^{jl}/\partial\beta^k$ exists. It follows that $\partial f^i/\partial\beta^k$ exists for (x) in the open set V and is of class C^{r-1} in (x) and these additional derivatives are continuous in $(x;\,\beta)$.

We see therefore that the existence of $\partial f^i/\partial\beta^k$ at $P_0, \cdots, P_n$ implies its existence in an open set of M. Let us show that this extends to all of M. Suppose we have proved the differentiability with respect to the β's of the functions representing the transformation (β) at all points P in an open set $A \subset M$. Suppose A has a boundary point P. As above, let $Q = T_0(P)$, U a neighborhood of Q, V a neighborhood of P so that $T(\bar{V}) \subset U$ for $T \in W$. We then choose

[8] The first proof of this for the analytic case was given us by Professor Marston Morse. Our proof has much in common with his.

the points $P_0, P_1, \cdots, P_n$ in the intersection $A \cdot V$ and reason as before that $f^i(x; \beta)$ is differentiable in (β) for (x) in an open set about P. There is a largest open set of M in which the transformation (β) is differentiable in (β) (i.e. the sum of all open sets in which this holds). This set is non-vacuous and we have just seen that it has no boundary point. It is therefore the whole of M. Thus we have

LEMMA 8. *If the functions determining the transformation $T \in W$ of M are differentiable in (β) at a set of points $P_0, P_1, \cdots, P_n$ as in Theorem 3, then they are differentiable in (β) at every point of M. Furthermore these derivatives are of class C^{r-1} (analytic) in (x) and these x-derivatives are continuous in $(x; \beta)$.*

We are now prepared to prove

THEOREM 9. *The parameters introduced into H in §7 make of H a Lie group.*

Suppose A_0, B_0 are isometries in H and $C_0 = B_0A_0$. Let $(\alpha^1, \cdots, \alpha^r) = (\alpha)$, (β), (γ) be coördinates for neighborhoods of A_0, B_0, C_0 respectively as provided in §7. Then for A, B near to A_0, B_0 the coördinates of $C = BA$ are given by functions $\gamma^i = \phi^i(\alpha; \beta)$ which we must prove to be of class C^1.

Under the imbedding of H in M^{n+1}, the isometry A_0 corresponds to the point $A_0(\overline{P})$. Let $(x) = (x^1, \cdots, x^m)(m = n(n+1))$ be coördinates for a neighborhood of $A_0(\overline{P})$; and let (y) be coördinates for a neighborhood of $C_0(\overline{P})$. Then the isometry B (near B_0) acting on M^{n+1} is given near $A_0(\overline{P})$ by functions $y^i = F^i(x; \beta)$. The functions $x^i = x^i(\alpha)$ defining the imbedding of the neighborhood (α) in M^{n+1} are of class C^1 with functional matrix of rank r. Similarly for the functions $y^i = y^i(\gamma)$ defining the imbedding of (γ) in M^{n+1}. Accordingly we can solve these latter for the γ's in terms of r of the y's: $\gamma^i = \gamma^i(y^{j_1}, \cdots, y^{j_r})$. We obtain in this way a representation of $\phi^i(\alpha; \beta)$:

$$\phi^i(\alpha; \beta) = \gamma^i(F^{j_1}(x(\alpha); \beta), \cdots, F^{j_r}(x(\alpha); \beta)). \tag{8.10}$$

By Lemma 7, $\partial F^j/\partial x^k$ exists and is continuous in $(x; \beta)$. Since γ^i and $x^i(\alpha)$ are of class C^1, it follows that $\partial\phi^i/\partial\alpha^k$ exists and is continuous in $(\alpha; \beta)$.

For the moment, let A_0 be the identity. Then the functions $F^j(\overline{P}; \beta)$ define the imbedding of the neighborhood (β) of B_0 in M^{n+1}. We know that these are of class C^1. Hence $\partial F^j/\partial\beta^k$ exists at the point $\overline{P}$. This means that at the points $P_0, \cdots, P_n$ of M the transformation (β) of M is differentiable in (β). By Lemma 8, this holds at every point of M. It follows that $\partial F^j/\partial\beta^k$ exists at every point of M^{n+1} and is continuous in $(x; \beta)$. It then follows from (8.10) that $\partial\phi^i/\partial\beta^k$ exists and is continuous in $(\alpha; \beta)$. This proves Theorem 9.

We have just seen that the functions $f^i(x; \beta)$ defining the transformation B of M have β-derivatives for all points (x) in M. By Lemma 8, if M is at least of class C^2, $\partial f^i/\partial\beta^k$ has x-derivatives continuous in $(x; \beta)$. This concludes the proof of our principal result.

THEOREM 10. *Any closed group of isometries of a Riemannian manifold of class C^r $(r \geqq 2)$ is a Lie group of isometries.*

9. REMARKS. We have seen that the parameters we have introduced in H make it a Lie group of class C^1. If the class of M is > 2, the question arises if

H is of a higher class. This reduces to the same question about a manifold of equivalent points. We have not settled this point.

The classical theory of motions concerns itself with Lie group germs of isometries in the neighborhood of a point of a Riemannian manifold. This suggests the question: *Is any locally compact group germ of local isometries a Lie group germ*? We hope to consider this in a later paper.

UNIVERSITY OF MICHIGAN, PRINCETON UNIVERSITY.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

ÜBER DIE AUSDRÜCKE DER GESAMTENERGIE UND DES GESAMTIMPULSES EINES MATERIELLEN SYSTEMS IN DER ALLGEMEINEN RELATIVITÄTSTHEORIE

VON PH. FREUD

(Received October 12, 1938)

Es bedarf einer umständlichen Überlegung, um in der allgemeinen Relativitätstheorie zu beweisen, dass der Gesamtimpuls und die Gesamtenergie (Energie + "Pseudoenergie") eines materiellen Systems unabhängig von allen Änderungen der Feldgrössen und des Koordinatensystems sind, die sich nicht auf die unmittelbare Nähe der Oberfläche erstrecken. Die Behauptung wird aber evident, wenn es gelingt, Impuls und Energie als *Flüsse durch die Oberfläche* darzustellen. Dies ist in dem Fall einer einzelnen Masse mit zentrischer Symmetrie von Weyl[1] geleistet worden. Für das Feld beliebig verteilter Materie soll es im folgenden geschehen; es gelingt, indem man die sechzehn Komponenten $\mathfrak{U}^i{}_k$ der Energie- und Impulsdichte und ihrer Strömungsdichten auf die einfache Form:

$$\mathfrak{U}^i{}_k = \frac{\partial \mathfrak{A}^{iv}{}_k}{\partial x^v} \tag{1}$$

bringt, in der $\mathfrak{A}^{in}{}_k$ in den Indizes i und n alterniert. Aus (1) folgt dann nämlich unmittelbar:

$$\frac{\partial \mathfrak{U}^\mu{}_k}{\partial x^\mu} = \frac{\partial^2 \mathfrak{A}^{\mu v}{}_k}{\partial x^v \partial x^\mu} = 0. \tag{2}$$

Wegen (1) lässt sich in dem Ausdruck des Gesamtimpulses und der Gesamtenergie eines Systems, das den Körper V mit den Volumelementen dv erfüllt:

$$\mathfrak{J}_k = \int_v \mathfrak{U}^4{}_k \, dv \tag{3}$$

(x_4 ist die Zeitkoordinate), die rechte Seite in ein Integral über die Oberfläche S mit den Flächenelementen $dS_{(a)} (a = 1, 2, 3)$ verwandeln:

$$\mathfrak{J}_k = \int_S \mathfrak{A}^{4\alpha}{}_k \, dS_{(\alpha)}, \tag{4}$$

womit die verlangte Darstellung erreicht und die Behauptung bewiesen ist.

[1] Weyl, *Raum, Zeit, Materie*, 5. Aufl., Berlin, 1923, S. 275.

Nun soll die Umformung (1) bewerkstelligt und damit der Ausdruck für $\mathfrak{A}^{in}{}_{k}$ ermittelt werden. Man bestimmt den zum Masstensor g_{ik} inversen Tensor aus:

$$g_{i\mu}g^{\mu k} = \delta_i^k \qquad (1 \text{ für } i = k,\ 0 \text{ für } i \neq k)$$

und bildet aus ihm und der Determinante $-g$ der g_{ik} die Grössen:

$$\mathfrak{g}^{ik} = \sqrt{g}\, g^{ik}.$$

Aus den Dreizeigersymbolen:

$$\Gamma^i_{kl} = \tfrac{1}{2} g^{i\mu}\left(\frac{\partial g_{\mu k}}{\partial x^l} + \frac{\partial g_{\mu l}}{\partial x^k} - \frac{\partial g_{kl}}{\partial x^\mu}\right)$$

bildet man den Krümmungstensor:

$$R^i{}_{klm} = \frac{\partial \Gamma^i_{kl}}{\partial x^m} - \frac{\partial \Gamma^i_{km}}{\partial x^l} + \Gamma^i_{\mu m}\Gamma^\mu_{kl} - \Gamma^i_{\mu l}\Gamma^\mu_{km} \tag{5}$$

und die Grössen:

$$\mathfrak{R}^i{}_k = \mathfrak{g}^{i\nu} R^\rho{}_{k\rho\nu}, \tag{6}$$

$$\mathfrak{G} = \mathfrak{g}^{\mu\nu}(\Gamma^\rho_{\mu\sigma}\Gamma^\sigma_{\rho\nu} - \Gamma^\rho_{\mu\nu}\Gamma^\sigma_{\rho\sigma}). \tag{7}$$

Die $\mathfrak{U}^i{}_k$ sind definiert durch[2]:

$$2\mathfrak{U}^i{}_k = \delta^i{}_k(\mathfrak{R}^\mu{}_\mu + \mathfrak{G}) - 2\mathfrak{R}^i{}_k + \left(\Gamma^i_{\mu\nu}\frac{\partial \mathfrak{g}^{\mu\nu}}{\partial x^k} - \Gamma^\nu_{\mu\nu}\frac{\partial \mathfrak{g}^{\mu i}}{\partial x^k}\right). \tag{8}$$

Der erste Bestandteil der rechten Seite wird schon bei Weyl[3] in eine Divergenz verwandelt:

$$\delta^i_k(\mathfrak{R}^\mu{}_\mu + \mathfrak{G}) = \frac{\partial}{\partial x^\nu}(\mathfrak{g}^{\mu\nu}\Gamma^\rho_{\mu\rho} - \mathfrak{g}^{\mu\rho}\Gamma^\nu_{\mu\rho})\delta^i_k. \tag{9}$$

Dasselbe soll mit dem übrigen Teil geschehen:

Aus den Identitäten[4]:

$$\frac{\partial \mathfrak{g}^{mi}}{\partial x^n} = -\mathfrak{g}^{\rho i}\Gamma^m_{\rho n} - \mathfrak{g}^{\rho m}\Gamma^i_{\rho n} + \mathfrak{g}^{mi}\Gamma^\rho_{\rho n},$$

$$\frac{\partial \mathfrak{g}^{i\nu}}{\partial x^\nu} = -\mathfrak{g}^{\mu\nu}\Gamma^i_{\mu\nu}$$

folgt:

$$\Gamma^i_{\mu k}\frac{\partial \mathfrak{g}^{\mu\nu}}{\partial x^\nu} - \Gamma^\nu_{\mu k}\frac{\partial \mathfrak{g}^{\mu i}}{\partial x^\nu} = -\mathfrak{g}^{\mu i}(\Gamma^\rho_{\mu k}\Gamma^\nu_{\rho\nu} - \Gamma^\rho_{\mu\nu}\Gamma^\nu_{\rho k}) + \mathfrak{g}^{\mu\nu}(\Gamma^\rho_{\mu k}\Gamma^i_{\nu\rho} - \Gamma^\rho_{\mu\nu}\Gamma^i_{\rho k}). \tag{10}$$

---

[2] Vgl. etwa: Pauli, *Relativitätstheorie*, Lpz. 1921, S. 622, Gl. (184), (185) und S. 597, Gl. (109).

[3] Weyl, l.c., S. 274.

[4] Weyl, l.c., S. 129, Gl. (60′), (60″).

Wir benützen ferner die beiden evidenten Umformungen:

$$\Gamma^i_{\mu\nu}\frac{\partial \mathfrak{g}^{\mu\nu}}{\partial x^k} - \Gamma^\nu_{\mu\nu}\frac{\partial \mathfrak{g}^{\mu i}}{\partial x^k} = \frac{\partial}{\partial x^k}(\mathfrak{g}^{\mu\nu}\Gamma^i_{\mu\nu} - \mathfrak{g}^{\mu i}\Gamma^\nu_{\mu\nu}) - \left(\mathfrak{g}^{\mu\nu}\frac{\partial \Gamma^i_{\mu\nu}}{\partial x^k} - \mathfrak{g}^{\mu i}\frac{\partial \Gamma^\nu_{\mu\nu}}{\partial x^k}\right) \tag{11}$$

$$-\left(\mathfrak{g}^{\mu\nu}\frac{\partial \Gamma^i_{\mu k}}{\partial x^\nu} - \mathfrak{g}^{\mu i}\frac{\partial \Gamma^\nu_{\mu k}}{\partial x^\nu}\right) = -\frac{\partial}{\partial x^\nu}(\mathfrak{g}^{\mu\nu}\Gamma^i_{\mu k} - \mathfrak{g}^{\mu i}\Gamma^\nu_{\mu k}) + \left(\Gamma^i_{\mu k}\frac{\partial \mathfrak{g}^{\mu\nu}}{\partial x^\nu} - \Gamma^\nu_{\mu k}\frac{\partial \mathfrak{g}^{\mu i}}{\partial x^\nu}\right) \tag{12}$$

und folgern aus (5):

$$\begin{aligned}\mathfrak{g}^{\mu\nu}R^i_{\mu\nu k} - \mathfrak{g}^{\mu i}R^\nu_{\mu\nu k} &= \left(\mathfrak{g}^{\mu\nu}\frac{\partial \Gamma^i_{\mu\nu}}{\partial x^k} - \mathfrak{g}^{\mu i}\frac{\partial \Gamma^\nu_{\mu\nu}}{\partial x^k}\right) + \mathfrak{g}^{\mu\nu}(\Gamma^i_{\rho k}\Gamma^\rho_{\mu\nu} - \Gamma^i_{\rho\nu}\Gamma^\rho_{\mu k}) \\ &\quad - \left(\mathfrak{g}^{\mu\nu}\frac{\partial \Gamma^i_{\mu k}}{\partial x^\nu} - \mathfrak{g}^{\mu i}\frac{\partial \Gamma^\nu_{\mu k}}{\partial x^\nu}\right) - \mathfrak{g}^{\mu i}(\Gamma^\nu_{\rho k}\Gamma^\rho_{\mu\nu} - \Gamma^\nu_{\rho\nu}\Gamma^\rho_{\mu k})\,.\end{aligned} \tag{13}$$

Endlich multiplizieren wir die Identität[5]:

$$0 = g_{s\rho}R^\rho{}_{mnk} + g_{m\rho}R^\rho{}_{snk}$$

mit $\mathfrak{g}^{mn}g^{is}$ und erhalten:

$$0 = \mathfrak{g}^{\mu\nu}R^i_{\mu\nu k} + \mathfrak{g}^{\mu i}R^\nu_{\mu\nu k}\,. \tag{14}$$

Man setzt nun die Summe der linken Seiten der sieben Gleichungen (8) bis (14) gleich der Summe der rechten Seiten; so entsteht, wenn man $\mathfrak{R}^i_k$ aus (6) einführt:

$$2\mathfrak{U}^i_k = \frac{\partial}{\partial x^\nu}[\delta^i_k(\mathfrak{g}^{\mu\nu}\Gamma^\rho_{\mu\rho} - \mathfrak{g}^{\mu\rho}\Gamma^\nu_{\mu\rho}) + \delta^\nu_k(\mathfrak{g}^{\mu\rho}\Gamma^i_{\rho\mu} - \mathfrak{g}^{i\rho}\Gamma^\mu_{\rho\mu}) - (\mathfrak{g}^{\mu\nu}\Gamma^i_{\mu k} - \mathfrak{g}^{\mu i}\Gamma^\nu_{\mu k})].$$

Das ist die verlangte Darstellung (1), wenn man:

$$\mathfrak{A}^{in}{}_k = \tfrac{1}{2}\begin{vmatrix} \delta^i_k & \delta^n_k & \delta^\mu_k \\ \mathfrak{g}^{i\rho} & \mathfrak{g}^{n\rho} & \mathfrak{g}^{\mu\rho} \\ \Gamma^i_{\rho\mu} & \Gamma^n_{\rho\mu} & \Gamma^\mu_{\rho\mu} \end{vmatrix}$$

setzt. $\mathfrak{A}^{in}{}_k$ ist eine Affintensordichte, die, wie es verlangt wurde, in i und n alterniert.

[5] Pauli, l.c., S. 592, Gl. (91), (92).

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

SETS OF POSTULATES FOR BOOLEAN GROUPS[1]

BY B. A. BERNSTEIN

(Received August 13, 1938; revised January 3, 1939)

1. Object

In a Boolean algebra, the operation $\circ$ given by $a \circ b = ab' + a'b$ is an operation with respect to which the elements of the algebra form a group.[2] The group is an (additive) abelian group in which every element is its own inverse. An abelian group in which every element is its own inverse I call a *Boolean* group. The object of my paper is to give a number of sets of postulates for (non-trivial) Boolean groups in terms free from the usual symbolism of Boolean algebra.

2. List of postulates

The postulates constituting the various sets are taken from the following list of conditions on an undefined class K and a binary operation $\circ$. In Postulates 2–2^{iv}, 3, 7–7^{v}, 8, 8′ there is to be supplied the restriction: *whenever the elements involved and their indicated combinations are in* K. The list follows.

0. *K has at least two distinct elements.*

1. *$a \circ b$ is in K whenever a, b are in K.*

2. $a \circ (b \circ c) = (a \circ b) \circ c.$

2′. $a \circ (b \circ c) = b \circ (c \circ a).$

2″. $a \circ (b \circ c) = c \circ (b \circ a).$

2‴. $(a \circ b) \circ c = (a \circ c) \circ b.$

2^{iv}. $a \circ (b \circ c) = (a \circ c) \circ b.$

3. $a \circ b = b \circ a.$

4. *There is an element 0 in K such that $a \circ 0 = a$ for every element a in K.*

5. *If the element 0 of Postulate 4 exists, and is unique, then for every element a in K there is an element a' in K such that $a \circ a' = 0$.*

6. *If the elements 0, a' of Postulates 4, 5 exist, and are unique, then $a' = a$.*

7. $a = (b \circ b) \circ a.$

7′. $a = (b \circ a) \circ b.$

7″. $a = (a \circ b) \circ b.$

7‴. $a = b \circ (b \circ a).$

7^{iv}. $a = b \circ (a \circ b).$

7^{v}. $a = a \circ (b \circ b).$

---

[1] Presented to the American Mathematical Society, April 9, 1938.

[2] See my two papers: (I) "Operations with respect to which the elements of a Boolean algebra form a group," *Trans. Amer. Math Soc.*, vol. 26 (1924), pp. 171–175, and vol. 27 (1925) p. 600; (II) "A set of postulates for Boolean algebra involving the operation of complete disjunction," these *Annals*, vol. 37 (1936), pp. 317–325.

8. *If* $a \circ b = c$ *then* $b = c \circ a$.
8′. *If* $a \circ b = c$ *then* $b = a \circ c$.
9. *For any two elements* a, b *in* K *there is an element* x *in* K *such that* $a \circ x = b$.

3. The postulate sets

The various sets of postulates are given in the table below. For each set is furnished a proof of the mutual independence of the postulates of the set. The independence examples are taken from the arithmetic systems i–xii following.

i. $K = 0; \quad a \circ b = 0.$

ii. $K = 0,1; \quad a \circ b = a + b + 0/(ab + b + 1) \bmod 2.$

iii. $K = 0,1,2; \quad a \circ b = a^2(b^2 + 2b) + a(2b^2 + 2b + 1) + b \bmod 3.$

iv. $K = 0,1; \quad a \circ b = a.$

v. $K = 0,1; \quad a \circ b = 0.$

vi. $K = 0,1; \quad a \circ b = ab + a + b \bmod 2.$

vii. $K = 0,1,2; \quad a \circ b = a + b \bmod 3.$

viii. $K = 0,1; \quad a \circ b = 0 \div 0.$

ix. $K = 0,1; \quad a \circ b = b.$

x. $K = 0,1,2; \quad a \circ b = 2(a + b) \bmod 3.$

xi. $K = 0,1,2; \quad a \circ b = a + 2b \bmod 3.$

xii. $K = 0,1; \quad a \circ b = a \div b.$

The table of postulate sets follows. In this table the independence system for the i^{th} postulate in any set occupies the i^{th} place among the independence systems for that set. Thus, for Set II, whose postulates are 0, 1, 2′, 7, the independence systems are respectively i, viii, ix, v.

TABLE OF POSTULATE SETS

Set	Postulates	Independence Systems
I	0, 1, 2, 3, 4, 5, 6	i, ii, iii, iv, v, vi, vii
II	0, 1, 2′, 7	i, viii, ix, v
III	0, 1, 2′, 7′	i, viii, x, v
IV	0, 1, 2′, 7″	i, viii, iv, v
V	0, 1, 2′, 7‴	i, viii, ix, v
VI	0, 1, 2′, 7^{iv}	i, viii, x, v
VII	0, 1, 2′, 7^{v}	i, viii, iv, v
VIII	0, 1, 2, 7^{iv}	i, viii, x, v
IX	0, 1, 2″, 7	i, viii, ix, v
X	0, 1, 2‴, 7	i, viii, ix, v
XI	0, 1, 2^{iv}, 7	i, viii, ix, iv

XII	0, 1, 2, 8	i, viii, x, ix
XIII	0, 1, 2′, 8	i, viii, x, vii
XIV	0, 2, 8, 9	i, x, vii, viii
XV	0, 2′, 8, 9	i, x, vii, viii
XVI	0, 2^{iv}, 8, 9	i, x, vii, viii
XVII	0, 1, 2, 3, 7‴	i, viii, x, ix, v
XVIII	0, 2, 3, 9, 8′	i, x, ix, viii, vii
XIX	0, 1, 2″, 7‴, 3	i, viii, x, v, xi
XX	0, 2″, 9, 3, 8′	i, x, xii, xi, vii

4. Remarks on the postulate sets

All the sets in the above table, except Set I, are free from unconditioned existence postulates, other than Postulate 0, which merely rules out trivial systems.

Set I gives a Boolean group directly in accordance with the definition of a Boolean group. It contains both the ordinary associative law and the commutative law. Postulates 0–5 of this set, taken by themselves, form a set of postulates for additive abelian groups.

Sets II–VII are briefer than Set I. In each of these sets the *cyclic* associative law is used instead of the ordinary associative law; and each set contains the closure condition and a (direct) *cancellation* law, but no commutative law.

Sets VIII–XI are variations of Sets II–VII. In these sets, other associative laws than the cyclic associative law are used.

Sets XII and XIII have a *transposition* law, instead of a cancellation law.

Sets XIV–XVI contain no closure postulates.

Set XVII is a set, briefer than Set I, in which the ordinary associative law and the commutative law both appear.

Set XVIII, like Set I, "embeds" a set of postulates for additive abelian groups (Postulates 0, 2, 3, 9); but it is briefer than Set I.

Set XIX embeds a set of postulates for subtractive abelian groups (Postulates 0, 1, 2″, 7‴).[3]

Set XX embeds both a set of postulates for additive abelian groups (Postulates 0, 2″, 9, 3) and a set of postulates for subtractive abelian groups (Postulates 0, 2″, 9, 8′).

[3] See my "Postulates for abelian groups and fields in terms of non-associative operations," *Trans. Amer. Math. Soc.*, vol. 43 (1938), pp. 1–6.

THE UNIVERSITY OF CALIFORNIA.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

SUBFIELDS AND AUTOMORPHISM GROUPS OF p-ADIC FIELDS[1]

BY SAUNDERS MAC LANE

(Received October 27, 1938)

1. Introduction

The fundamental theorem on the structure of perfect fields states that if a field K of given characteristic is perfect (or topologically complete) with respect to a discrete non-archimedean valuation $|\ a\ |$, with the properties,

$$(1) \qquad |\ ab\ | = |\ a\ |\ |\ b\ |, \qquad |\ a + b\ | \leqq \text{Max}\ (|\ a\ |, |\ b\ |),$$

then K is uniquely determined by its field $\mathfrak{R}$ of residue classes, except in the case when K has characteristic 0, $\mathfrak{R}$ has characteristic p, and K is "ramified" over its rational subfield. The first proof of this structure theorem in the general case is due to Hasse and Schmidt [2], but for an imperfect residue class field $\mathfrak{R}$ their proof unfortunately uses an elaborate and unproven[2] lemma on the generation of $\mathfrak{R}$ by a Steinitz field tower. A second proof of the structure theorem has been given by Witt [13] and Teichmüller [10] and [11a]. For the case of an (algebraically) perfect residue class field $\mathfrak{R}$ Witt's proof uses a sophisticated "vector" analysis construction of the p-adic fields K, and Teichmüller's methods then treat the imperfect fields K on this basis.

The structure theorem involves two steps: first, the construction of a discrete complete field K with a given characteristic and a given residue class field $\mathfrak{R}$; second, the demonstration that the so constructed field is unique. The separation of these two steps shows that the previous constructions of K have been needlessly involved and can be replaced by an elementary stepwise construction (§3). This construction can then be so combined with the methods of Teichmüller for the imperfect fields $\mathfrak{R}$ as to give an elementary proof of the structure theorem (Corollary to Theorem 8 in §6). The cases when K and its residue class field $\mathfrak{R}$ have the same characteristic can be reduced simply by known methods (see Teichmüller [11a]), as K is then a power series field. Hence we consider only the so-called p-adic and $\mathfrak{p}$-adic fields; that is, discrete complete fields of characteristic zero with residue class fields of characteristic p.

Problems of the relative structure of complete fields can also be treated by these methods. One such question, which is suggested by investigations of valuations of higher rank (Mac Lane [4]) is this: Given two unramified p-adic extensions K and K' of a given p-adic field k with residue class field $\mathfrak{k}$, such that

[1] Presented to the American Mathematical Society, Sept. 10, 1938.

[2] Difficulties present in the proof of such a Lemma are discussed in another paper by the author, entitled "Steinitz field towers for modular fields," forthcoming in the Transactions of the Am. Math. Soc.

K and K' have the same field $\mathfrak{R}$ of residue classes, are K and K' analytically equivalent over k? If $\mathfrak{k}$ and $\mathfrak{R}$ are finite fields, then k is an ordinary field of p-adic numbers, and the answer is known to be yes: K and K' are equivalent. For infinite fields $\mathfrak{R}/\mathfrak{k}$, any two extensions K and K' are still equivalent, as stated in Theorem 8 of §6, under certain restriction on the extension $\mathfrak{R}/\mathfrak{k}$. That some such restriction is necessary is indicated in §6 by an example.

Preliminary instances of this relative structure theory include the case of (transcendentally) separable extensions $\mathfrak{R}/\mathfrak{k}$, in §3, and the case of perfect fields $\mathfrak{R}$ and $\mathfrak{k}$, in §4. Two essential tools are the notions of multiplicative representatives (§2), and p-bases of imperfect fields, both due to Teichmüller.

Another question of relative structure of a p-adic field K concerns the existence of a p-adic subfield or intermediate field k with a specified residue class subfield $\mathfrak{k}$. The results on this problem given in §7 include the theorem that there is exactly one such field k if $\mathfrak{k}$ is perfect.

The group G of all automorphisms of a p-adic field K has an inertial subgroup G_1, consisting of automorphisms which leave all residue classes fixed, as well as certain ramification subgroups G_n. The quotient group G/G_1 is isomorphic (§8) to the automorphism group of the residue class field $\mathfrak{R}$, while each quotient group G_n/G_{n+1} is abelian and can be described explicitly in terms of $\mathfrak{R}$ (see §9). The results depend on certain constructions of automorphisms which are obtained from the relative structure theory of the preceding sections. The capstone of these investigations is the theorem of §10 which states that the subfield of elements left fixed by every automorphism of the inertial group G_1 is exactly that subfield k whose residue class field is the maximal perfect subfield of the whole residue class field. Strangely enough, the field of elements left fixed by the automorphisms of any ramification group G_n is the same field k!

2. Multiplicative representatives

A rank one "exponential" *valuation*[3] V of a field K is a real valued function $V(a)$ defined for $a \neq 0$ in K and having for all a and b the properties

$$V(ab) = V(a) + V(b), \qquad V(a+b) \geqq \operatorname{Min}(Va, Vb). \tag{2-1}$$

Associated with such a valuation there is the *value group* Γ, composed of all numbers $V(a)$, and the *residue class field* (res. field) $\mathfrak{R} = B/P$, obtained from the valuation ring B, which consists of all elements a with $Va \geqq 0$, and the prime ideal P, which contains all elements a with $Va > 0$. Denote by $H(a)$ the residue class of a in $\mathfrak{R}$, if a is in B, and set $H(a) = \infty$ if a is not in B. Then $H(a)$ is in the set $\{\mathfrak{R}, \infty\}$ composed of the elements of $\mathfrak{R}$ and the symbol ∞, while the many-one correspondence $a \to H(a)$ of K to $\{\mathfrak{R}, \infty\}$ has the properties that $Ha \neq \infty$ and $Hb \neq \infty$ imply

$$H(a+b) = H(a) + H(b), \qquad H(ab) = (Ha)(Hb). \tag{2-2}$$

[3] The non-archimedean norm $|a|$ of (1) in the introduction generates an exponential valuation $V(a) = -\log|a|$.

The functions H and V are connected by the logical equivalence of the three statements

$$H(a) = 0 \leftrightarrow H(1/a) = \infty \leftrightarrow Va > 0. \tag{2-3}$$

Such a homomorphism H reciprocally determines a corresponding V. By a *valuation* of K we henceforth mean a pair of functions $\{V, H\}$ consisting of a valuation function V and a homomorphism H of K on some $\{\mathfrak{R}, \infty\}$ with the properties (1), (2), and (3). Since V determines H and conversely, this simultaneous use of V and H is more symmetric; it also avoids conventions as to res. fields.[4] For instance if $K \subset L$, a valuation $\{W, H'\}$ of L is an *extension* of a valuation $\{V, H\}$ of K, if the functions W and H' are respectively extensions of V and H. Necessarily, then, the res. field $\mathfrak{L}$ of L contains that of K and the value group of L contains that of K. The extension L/K is *unramified* if these value groups are actually the same.

A valuation V is *discrete* if its value group Γ is isomorphic to the group of rational integers. A field K is *complete* in a valuation $\{V, H\}$, if it is complete in the topology introduced by the norm $|a| = e^{-Va}$. A discrete valuation of a field K is called $\mathfrak{p}$-*adic* if K is complete and of a characteristic 0 and if the res. field $\mathfrak{R}$ has as characteristic a prime p. A *prime element* π of K then is an element of minimum positive value in K. A $\mathfrak{p}$-adic valuation is called p-adic if the rational prime p is a prime element.

A *map* or *analytic isomorphism* T of K in K', where K and K' are fields with respective valuations $\{V, H\}$ and $\{V', H'\}$, is defined as an isomorphism $a \leftrightarrow a^T$ of K to a subset of K' which preserves valuations: $V'(a^T) = V(a)$. If the images a^T include all the elements of K, then T is an isomorphism of K *on* K'. Any isomorphism T generates a corresponding isomorphism τ of the res. field $\mathfrak{R}$ in a subfield of the res. field $\mathfrak{R}'$ of K', by the rule

$$[H(a)]^\tau = H'(a^T). \tag{2-4}$$

The res. field $\mathfrak{R}$ may be replaced by this isomorphic subfield $\mathfrak{R}^\tau$ of $\mathfrak{R}'$ without changing the function V in the given valuation $\{V, H\}$. Hence, given such an isomorphism, we shall often assume that $\mathfrak{R} \subset \mathfrak{R}'$ and that T leaves the res. field $\mathfrak{R}$ elementwise fixed, that is

$$H(a) = H'(a^T) \qquad (a \text{ in } K). \tag{2-5}$$

The congruence $a \equiv b \pmod{\pi^n}$, for π a prime element of a $\mathfrak{p}$-adic field K and for a and b in the valuation ring of K, means that $V(a - b) \geqq nV(\pi)$. In other words $a \equiv b \pmod{\pi^n}$ implies $a = b + d$ for some d with $V(d) \geqq V(\pi^n)$. Therefore

$$a^p = (b + d)^p = b^p + pb^{p-1}d + \cdots + pbd^{p-1} + d^p \equiv b^p \pmod{\pi^{n+1}},$$

hence the conclusion (Teichmüller [10], Lemma 8).

---

[4] Compare also the simpler proof obtained in Theorem 13 in §8.

LEMMA 1. *For elements a and b of a $\mathfrak{p}$-adic field, with $V(a) \geqq 0$, $V(b) \geqq 0$, $a \equiv b \pmod{\pi^n}$ implies $a^{p^m} \equiv b^{p^m} \pmod{\pi^{n+m}}$ for any $m \geqq 0$.*

For p-adic fields with perfect res. fields Teichmüller ([10], Lemma 5) has found a system R of "multiplicative representatives" of the residue classes. For any α in the res. field there is in R one *representative* a so that $H(a) = \alpha$.

LEMMA 2. *If $\mathfrak{k}$ is a perfect field of characteristic p, then any $\mathfrak{p}$-adic field K with residue class field $\mathfrak{K} \supset \mathfrak{k}$ contains a system R of representatives of the residue classes of $\mathfrak{k}$ which is uniquely determined by any one of the following three equivalent properties:*

(i) If a_n of K is in the residue class $\alpha^{p^{-n}}$ for α in $\mathfrak{k}$, then $\lim\ (n \to \infty) a_n^{p^n}$ exists and is the R-representative of α;

(ii) If b_n is the R-representative of $\beta^{p^{-n}}$, for β in $\mathfrak{k}$, then $b_{n+1}^p = b_n$.

(iii) If a, b, and c are R-representatives, respectively, of elements α, β, and γ of $\mathfrak{k}$, then $\alpha\beta = \gamma$ implies $ab = c$.

The proof of this Lemma is exactly like that given by Teichmüller for the special case $\mathfrak{K} = \mathfrak{k}$, and depends on using the description of (i) to define the system R.

I. RELATIVE STRUCTURE OF p-ADIC FIELDS

The Equivalence problem

Our primary problem is the relative equivalence of the extensions of complete fields with specified residue class extensions. Given a complete field k with a residue class field $\mathfrak{k}$ and two unramified complete extensions K and K' with the same residue class field $\mathfrak{K} = \mathfrak{K}'$; that is, given

(*) $\qquad K \supset k, \qquad K' \supset k$ complete, unramified with $\mathfrak{K} = \mathfrak{K}' \supset \mathfrak{k}$;

when is K analytically equivalent to K'? Here an *analytic equivalence* of K to K' means an analytic isomorphism T of K on K' which is the identity on k and also on the common residue class field $\mathfrak{K}$. The "discrete," or "p-adic" equivalence problem will always refer to the problem (*), in which k is assumed discrete or p-adic, as the case may be. A gothic $\mathfrak{k}$ or $\mathfrak{L}$ will always denote the residue class field of the field represented by the corresponding roman letter k or L, etc.

3. Extension of maps

Though the construction of p-adic fields with specified res. fields has usually been accomplished by elaborate methods, an elementary procedure ([4], §5, especially Theorem 4) gives a step by step construction. For instance, if a given p-adic field k with valuation $\{V, H\}$ has the res. field $\mathfrak{k}$, and if $\mathfrak{K} = \mathfrak{k}(\xi)$ is a given transcendental extension of $\mathfrak{k}$, then a corresponding extension of the complete field k can be found by constructing a transcendental extension $k(x)$ and defining V' and H' for any polynomial $f(x) = a_n x^n + \cdots + a_0$ in $k[x]$ by

$$(3\text{-}1) \qquad V'\left(\sum_{i=0}^{n} a_i x^i\right) = \operatorname*{Min}_i (Va_i), \qquad H'\left(\sum_{i=0}^{n} a_i x^i\right) = \sum_{i=0}^{n} (Ha_i)\xi^i.$$

The complete closure of $k(x)$ then yields a p-adic field with the res. field $\mathfrak{R} = \mathfrak{k}(\xi)$. Combining this with a similar treatment of algebraic extensions we have:

THEOREM 1. *Extension Theorem. Let k be a complete field with valuation $\{V, H\}$ and residue class field $\mathfrak{k}$, while $\mathfrak{R}$ is any extension of $\mathfrak{k}$. Then there exists an extension K of k, complete and unramified with respect to an extended valuation and having the residue class field $\mathfrak{R}$.*

For any prime res. field there is a corresponding field of p-adic numbers; hence the existence Theorem:

THEOREM 2. *Existence Theorem. There is a p-adic field K with any given residue class field $\mathfrak{R}$ of characteristic p.*

The extension Theorem 1 and subsequent studies of the uniqueness of this extension are closely related to the question of unramified division algebras with specified residue class division algebras, as treated by Witt ([12]) and Nakayama ([6]).

The problem of extending a given analytic isomorphism to parallel a specific residue class extension $\mathfrak{R}/\mathfrak{k}$ can be treated if $\mathfrak{R}$ has a separating transcendence basis[5] over $\mathfrak{k}$:

THEOREM 3. *Separable map extensions. Let two complete fields k and k' with the same residue class fields $\mathfrak{k}$ be analytically isomorphic under the map T which is the identity on $\mathfrak{k}$, and let K and K' be complete unramified extensions, $K \supset k$, $K' \supset k'$, with respective residue class fields $\mathfrak{R} \subset \mathfrak{R}'$. If the field $\mathfrak{R}$ has a separating transcendence basis $\mathfrak{y}$ over $\mathfrak{k}$, the given map T can be extended to an analytic isomorphism T^* of K on a subfield of K', which is again the identity on $\mathfrak{R}$. There is only one such extension T^* if and only if $\mathfrak{R}$ is algebraic over $\mathfrak{k}$.*

The most natural instance of this Theorem arises when K and K' have the same res. field $\mathfrak{R} = \mathfrak{R}'$, so that T^* maps K on all K'. The proof of this instance is the same as for the general case, and the more general case is useful later in §7.

PROOF. Because $\mathfrak{y}$ is a sep. trans. basis, $\mathfrak{R}/\mathfrak{k}$ can be built up by successive transcendental and sep. algebraic adjunctions. The required extension can then be found by transfinite induction if the case of a single transcendental or a separable algebraic extension can be treated (cf. MacLane [4], Lemma 2). *Transcendental case.* If $\mathfrak{R} = \mathfrak{k}(\xi)$ for ξ transcendental over $\mathfrak{k}$, and if x and x' are the elements from K and K' respectively representing ξ, then x and x' are transcendental over k and k', and both $k(x)$ and $k'(x')$ have the res. field $\mathfrak{k}(\xi)$. T can be extended in one and only one way to a value-preserving isomorphism T^* of $k(x)$ onto $k'(x')$ such that $x^{T^*} = x'$. Specifically, if $f(x) = a_n x^n + \cdots + a_0$ is a polynomial in $k[x]$, then $V(f)$ and $H(f)$ are given by (3-1) and $[f(x)]^{T^*} = \sum a_i^T x'^i$ is the desired extension. This T^*, extended in the natural way[6] to the complete closure of $k(x)$, is still the identity on $\mathfrak{R}$. *Separable algebraic case.* If $\mathfrak{R} = \mathfrak{k}(\beta)$, where β satisfies the irreducible, separable, and monic equation

---

[5] A subset $\mathfrak{y}$ of $\mathfrak{R}$ is a separating transcendence basis (sep. trans. basis) for $\mathfrak{R}$ over $\mathfrak{k}$ if the elements of $\mathfrak{y}$ are algebraically independent over $\mathfrak{k}$ and if $\mathfrak{R}$ is separable and algebraic over $\mathfrak{k}(\mathfrak{y})$.

[6] Hasse-Schmidt, [2], Fortsetzbarkeitsatz p. 26 and Satz 2 on p. 28.

$g(z) = 0$ over $\mathfrak{k}$, choose in $k[z]$ any monic polynomial $G(z)$ with the same degree as $g(z)$ and with coefficients representative of the coefficients of $g(z)$. Denote by $G'(z)$ the corresponding polynomial in K'. By the Hensel-Rychlik reducibility Theorem,[7] K contains exactly one root b of $G(z) = 0$ with residue class β. Furthermore $k(b)$ has res. field $\mathfrak{K} = \mathfrak{k}(\beta)$, and the map T can be extended to $k(b)$ by taking as b^T the correspondingly unique root b' of $G'(z) = 0$ with residue class β. This is clearly the only extension of T to $K = k(b)$, so that the Theorem is proven.

In the special case when $\mathfrak{K} = \mathfrak{K}'$ the extension T^* found by combining the above methods maps K on all K'. For, the map K^{T^*} is a subfield of K' with the same value group and the same res. field as K', hence is necessarily equal to K' (MacLane [4], Theorem 8). Hence a first equivalence Theorem, which applies without any assumption of discreteness:

THEOREM 4. *Separable Equivalence. In an unramified equivalence problem,*

$$(*) \qquad K \supset k, \qquad K' \supset k, \qquad \text{with } \mathfrak{K} = \mathfrak{K}' \supset \mathfrak{k},$$

the complete fields K and K' are analytically equivalent over k if the residue class field $\mathfrak{K}$ has a separating transcendence basis over the original residue class field $\mathfrak{k}$.

4. Complete fields with perfect residue class fields

THEOREM 5. *Given a discrete and unramified equivalence problem (*) with perfect residue class fields $\mathfrak{K}$ and $\mathfrak{k}$, the extensions K and K' are always analytically equivalent over k. More generally, if the given unramified extensions K and K' have distinct residue class fields $\mathfrak{K}$ and $\mathfrak{K}'$ with $\mathfrak{K} \subset \mathfrak{K}'$ and if $\mathfrak{K}$ and $\mathfrak{k}$ are perfect, there is an analytic isomorphism T of K in K' which leaves elements of k and residue classes of $\mathfrak{K}$ fixed.*

This Theorem, when proven, will have the following consequences:

COROLLARY 1. *If K and K' are p-adic fields with respective residue class fields $\mathfrak{K} \subset \mathfrak{K}'$, where $\mathfrak{K}$ is perfect, there is an analytic isomorphism of K into a subfield of K' which leaves fixed the residue classes of $\mathfrak{K}$.*

PROOF. Apply Theorem 5, using for k the minimal complete subfield (of p-adic numbers) contained in both K and K'. A special case of this corollary is the usual "structure theorem" for a perfect residue class field:

COROLLARY 2. *Any two p-adic fields with the same perfect residue class field are analytically isomorphic under a map which leaves each residue class fixed.*

Theorem 5 may be proven directly from the separable map extension of Theorem 3 if the res. field $\mathfrak{K}$ has characteristic 0. Suppose then that $\mathfrak{K}$ has characteristic p, and let the set $\mathfrak{Y}$ of elements η form a transcendence basis for $\mathfrak{K}/\mathfrak{k}$. The perfect field $\mathfrak{K}$ must contain all roots $\eta^{p^{-e}}$ and hence also the subfield $\mathfrak{K}_0 = \mathfrak{k}(\mathfrak{Y}^{p^{-\infty}})$ consisting of all rational functions of all quantities $\eta^{p^{-e}}$, for which η is in $\mathfrak{Y}$ and e is an integer. For each such η Lemma 2 provides multi-

[7] Albert, [1], Lemma on p. 296: Hasse-Schmidt, [2], p. 31. The proof given by Albert for the discrete case holds equally well for a non-discrete valuation.

plicative representatives y and y' in K and K' respectively. The set Y of all the representatives y in K is a set of independent transcendents over k because the homomorphic images $\mathfrak{Y}$ are algebraically independent over $\mathfrak{K}$. The multiplicative representative $y^{p^{-e}}$ of $\eta^{p^{-e}}$ is by Lemma 2, (ii) a uniquely determined p^{th} root of y. Therefore K contains the ring $R_0 = k[Y^{p^{-\infty}}]$, which consists of all polynomials f in the quantities $y^{p^{-e}}$, of the form

$$f(y_1, \cdots, y_m) = a_1(y_1^{q_{11}} \cdots y_m^{q_{1m}}) + \cdots + a_t(y_1^{q_{t1}} \cdots y_m^{q_{tm}}), \tag{4-1}$$

in which $y_1, \cdots, y_m$ are any elements of Y, each coefficient a_i is in k and each exponent q_{ij} is a rational fraction with denominator a power of p. The valuation $\{V, H\}$ of K has $Vy = 0$, $Hy = \eta$. The algebraic independence of the η's and the definition (2-2) of a homomorphism give, by computation in (4-1),

$$V(f) = \operatorname{Min}(Va_1, \cdots, Va_t) \tag{4-2}$$

$$H(f) = (Ha_1)(\eta_1^{q_{11}} \cdots \eta_m^{q_{1m}}) + \cdots + (Ha_t)(\eta_1^{q_{t1}} \cdots \eta_m^{q_{tm}}). \tag{4-3}$$

In other words, $H[f(y_1, \cdots, y_m)] = f^H(\eta_1, \cdots, \eta_m)$, where f^H denotes the polynomial obtained by applying H to each coefficient of f.

The set Y' of all multiplicative representatives y' of elements η in K' generates a similar ring $R_0' = k[Y'^{p^{-\infty}}] \subset K'$, for which the valuation and homomorphism are given by similar formulas. This means that the correspondence

$$T_0 : f(y_1, \cdots, y_m) \leftrightarrow f(y_1', \cdots, y_m')$$

is an analytic isomorphism of R_0 to R_0' which leaves fixed elements of k and of the residue-class ring $\mathfrak{k}[\mathfrak{Y}^{p^{-\infty}}]$. In natural fashion T_0 can be extended to the quotient field of R_0 and thence to the complete closure K_0 of this quotient field in K. The res. field of K_0 is $\mathfrak{K}_0 = \mathfrak{k}(\mathfrak{Y}^{p^{-\infty}})$; since $\mathfrak{Y}$ was chosen a trans. basis for $\mathfrak{K}/\mathfrak{k}$, the whole res. field is algebraic over the smaller field $\mathfrak{K}_0$. Furthermore $\mathfrak{K}_0$, like $\mathfrak{k}$, is perfect, so that $\mathfrak{K}$ must be separable over $\mathfrak{K}_0$. The map of K_0 in K' can therefore be extended by Theorem 3 to a map of K in K'. By construction, this map leaves fixed all elements of k and all res. classes.

5. Teichmüller's embedding process

Imperfect residue class fields may be analysed by p-bases. In any field $\mathfrak{K}$ of characteristic p consider the generation of $\mathfrak{K}$ over the subfield $\mathfrak{K}^p$ consisting of the p^{th} powers of all its elements. A subset $\mathfrak{M}$ of $\mathfrak{K}$ is *p-independent* if $\mathfrak{K}^p(\mathfrak{M}')$ is a proper subfield of $\mathfrak{K}^p(\mathfrak{M})$ whenever $\mathfrak{M}'$ is a proper subset of $\mathfrak{M}$. A subset $\mathfrak{M}$ is a *p-basis* of $\mathfrak{K}$ if $\mathfrak{M}$ is p-independent and if $\mathfrak{K}^p(\mathfrak{M}) = \mathfrak{K}$. Teichmüller ([11], §3), who gave definitions equivalent to these, showed that every $\mathfrak{K}$ has a p-basis and can be obtained from the subfield $\mathfrak{K}^{p^n}$ of the $p^{n\text{th}}$ powers ($n \geqq 0$) of all its elements by the adjunction of any p-basis $\mathfrak{M}$;

$$\mathfrak{K} = \mathfrak{K}^{p^n}(\mathfrak{M}). \tag{5-1}$$

A p-adic field with imperfect res. field can be embedded in a larger field

with perfect res. field by successive extensions like those of Theorem 1. Teichmüller's exposition of this process[8] may be summarized thus:

THEOREM 6. (Teichmüller.) *If the $\mathfrak{p}$-adic field K has the residue class field $\mathfrak{K}$ containing a p-independent subset $\mathfrak{M}$, and if $M \subset K$ is a fixed set of representatives in K of the residue classes of $\mathfrak{M}$, then there exists a complete unramified extension L of K with a residue class field $\mathfrak{L} = \mathfrak{K}(\mathfrak{M}^{p^{-\infty}})$, whose maximal perfect subfield contains $\mathfrak{M}$, such that the multiplicative representatives in L of each element of $\mathfrak{M}$ are exactly the corresponding elements of M. If $\mathfrak{M}$ is a p-basis of $\mathfrak{K}$, $\mathfrak{L}$ is the minimal perfect extension $\mathfrak{K}^{p^{-\infty}}$ of $\mathfrak{K}$.*

In this theorem $\mathfrak{K}(\mathfrak{M}^{p^{-\infty}})$ denotes the field obtained from $\mathfrak{K}$ by adjoining all elements $\mu^{p^{-e}}$, for e an integer and μ in $\mathfrak{M}$. If $\mathfrak{M}$ is not a p-basis, a p-basis for $\mathfrak{L}$ may be found by the following subsequently useful process (proof omitted).

LEMMA 3. *If the disjoint subsets $\mathfrak{M}$ and $\mathfrak{N}$ of a field $\mathfrak{K}$ together constitutes a p-basis of $\mathfrak{K}$, then $\mathfrak{N}$ is a p-basis of $\mathfrak{K}(\mathfrak{M}^{p^{-\infty}})$.*

The existence and uniqueness of certain subfields $K \subset L$ in Theorem 6 were also established by Teichmüller for the p-adic case:

THEOREM 7. *If the p-adic field L has a residue class field $\mathfrak{L}$ with the maximal perfect subfield $\mathfrak{L}^{p^\infty}$, if $\mathfrak{M}$ is a p-basis of a subfield $\mathfrak{K} \subset \mathfrak{L}^{p^\infty}$, and if M is the set of multiplicative representatives of $\mathfrak{M}$ in L, then there is one and only one p-adic subfield K of L which contains M and has the residue class field $\mathfrak{K}$.*

PROOF. The existence of K, proven by Teichmüller by a suitable appeal to the Witt vector addition, can be obtained in elementary fashion. First construct any p-adic field K' with res. field $\mathfrak{K}$ (Theorem 1), and let M' be any set of representatives of $\mathfrak{M}$ in K'. By Theorem 6 embed K' in a p-adic field L_0 with res. field $\mathfrak{L}_0 = \mathfrak{K}^{p^{-\infty}} \subset \mathfrak{L}$ in such a way that the elements of M' become multiplicative representatives. By Corollary 1 of Theorem 5 there is a map T of L_0 in the given L. The definition of multiplicative representatives in Lemma 2 is invariant and T leaves residue classes fixed, so that T carries the set M' into the multiplicative representatives M. Therefore the image of K' under T is the required subfield K of L.

The uniqueness of K as asserted in Theorem 7 will be established in a more general form, useful subsequently in Theorem 12 and Lemma 6:

LEMMA 4. *In a $\mathfrak{p}$-adic field L any two complete subfields K and K' with the same residue class field $\mathfrak{K}$ are identical ($K = K'$) if K and K' contain in common some prime element π of L and a set M of representatives for some p-basis $\mathfrak{M}$ of the residue class field $\mathfrak{K}$.*

Note that $\mathfrak{L}$ need not be perfect nor M multiplicative, while if L is p-adic, the prime element p of L is necessarily present in every K, so the hypothesis on π is true.

Our proof arises from a simplification of certain res. class formulas due to

---

[8] Teichmüller, [10], pp. 146–147, Lemmas 14 to 18. This part of the paper uses no vector calculus, and proves Theorem 6 below only for $\mathfrak{M}$ a p-basis. Our more general statement is established by exactly the same methods.

Teichmüller ([10], Lemma 17). For any given b' of non-negative value in K' we need only find elements b_n in K such that $b' \equiv b_n \pmod{\pi^n}$ for each n. For then $b' = \lim b_n$ is in the complete field K, $K' \subset K$, and hence $K' = K$.

For $n = 1$, $b_1 \equiv b'$ exists because K and K' have the same res. field. In general, b_{n+1} is to be constructed by induction from b_n. As in (5-1), the res. class β of b' can be expressed as a polynomial $f(\mathfrak{M})$ in elements of the p-basis $\mathfrak{M}$ with coefficients α^{p^n} in $\mathfrak{R}^{p^n}$. If each coefficient α is replaced by a representative a' in K' and if each variable ξ of $\mathfrak{M}$ is replaced by its representative x in the set M, there results a polynomial $F'(M)$ lying in the residue class $f(\mathfrak{M})$, so that $F'(M) \equiv b' \pmod{\pi}$. Therefore $b' = F'(M) + c'\pi$ for some c' in K' with $Vc' \geqq 0$. Construct $F(M)$ from $F'(M)$ by replacing each coefficient a'^{p^n} by a^{p^n}, where a in K is chosen so that $a \equiv a' \pmod{\pi}$. Then by Lemma 1, $F'(M) \equiv F(M) \pmod{\pi^{n+1}}$, while the induction assumption yields a c_n in K with $c' \equiv c_n \pmod{\pi^n}$. Consequently $b' = F'(M) + c'\pi \equiv F(M) + c_n\pi \pmod{\pi^{n+1}}$, so that $F(M) + c_n\pi$ is the required element b_{n+1} of K.

An analogous proof yields a related Lemma concerned with maps, to be used in Lemma 6 of §8:

LEMMA 5. *Let K/k be an unramified $\mathfrak{p}$-adic extension with residue class fields $\mathfrak{R}/\mathfrak{k}$, and let $\mathfrak{m} \subset k$ be a system of representatives for a p-basis $\mathfrak{m}$ of $\mathfrak{k}$. Then any analytic automorphism T of K which carries $\mathfrak{m}$ into itself and leaves $\mathfrak{k}$ and a prime element π of k elementwise fixed must leave k elementwise fixed.*

This includes the useful case when $K = k$, while if K is p-adic the hypothesis on π is superfluous. To prove the Lemma, we need only show $b^T \equiv b \pmod{\pi^n}$ for each n and each b in k with $Vb \geqq 0$. For $n = 1$ this must hold, since by hypothesis T leaves the res. class of b fixed. In the induction proof, $b = F(m) + c\pi$, as before, so that $b^T = F^T(m) + c^T\pi$, where $c^T \equiv c \pmod{\pi^n}$ and F^T is obtained from F by replacing each coefficient a^{p^n} by $(a^T)^{p^n}$. But

$$a \equiv a^T \pmod{\pi}, \qquad a^{p^n} \equiv (a^T)^{p^n} \pmod{\pi^{n+1}},$$

as in Lemma 1, so that we have the required conclusion $b^T \equiv b \pmod{\pi^{n+1}}$.

6. Equivalence of unramified extensions

A more general result than the equivalence Theorems 4 and 5 is given by

THEOREM 8. *Given a p-adic unramified equivalence problem (*) of Theorem 4, the extensions K and K' are analytically equivalent over k if some p-basis $\mathfrak{m}$ of the original residue class field $\mathfrak{k}$ is p-independent in the extended field $\mathfrak{R}$. If k is not p-adic but $\mathfrak{p}$-adic, the theorem is still true, provided k is unramified over some $\mathfrak{p}$-adic subfield k_0 whose residue class field is perfect.*

First some remarks: Any two p-adic fields K and K' with the same res. field $\mathfrak{R}$ both contain the (prime) field k of p-adic numbers. The res. field $\mathfrak{k}$ of k is finite and perfect, hence has a void p-basis, so that Theorem 8 includes the usual structure theorem for arbitrary res. fields:

COROLLARY 1. *Any two p-adic fields with the same residue class field are analytically isomorphic under a map which preserves residue classes.*

The essential hypothesis of Theorem 8 on $\mathfrak{K}/\mathfrak{k}$ is the requirement that some p-basis of $\mathfrak{k}$ is a subset of a p-basis of $\mathfrak{K}$. It can be shown that this will be the case if and only if every subset of $\mathfrak{k}$ p-independent in $\mathfrak{k}$ remains p-independent in $\mathfrak{K}$. An extension $\mathfrak{K}/\mathfrak{k}$ with this property is said to *preserve p-independence*. An algebraic extension $\mathfrak{K}/\mathfrak{k}$ preserves p-independence if and only if it is separable. The force of this hypothesis may also be indicated by the following examples of extensions $\mathfrak{K}/\mathfrak{k}$ preserving p-independence:

(i) Any extension $\mathfrak{K}$ of a perfect field $\mathfrak{k}$;

(ii) Any separable algebraic extension $\mathfrak{K}$ of any field $\mathfrak{k}$;

(iii) Any purely transcendental extension $\mathfrak{K}$ of any field $\mathfrak{k}$;

(iv) Any extension $\mathfrak{K}$ with a separating transcendence basis over $\mathfrak{k}$ (see §3).

The p-bases necessary to prove that these extensions preserve independence may be found directly or by the constructions for p-bases given by Teichmüller in [11]. These examples do not exhaust the independence-preserving extensions; for instance, over $\mathfrak{k} = \mathfrak{P}(x)$, $\mathfrak{K} = \mathfrak{P}(x, t, t^{p^{-1}}, t^{p^{-2}}, \cdots)$ preserves p-independence if $\mathfrak{P}$ is perfect and x and t are algebraically independent over $\mathfrak{P}$.

Proof of Theorem 8. The given p-basis $\mathfrak{m}$ of $\mathfrak{k}$ is p-independent in $\mathfrak{K}$ and hence can be enlarged to a p-basis $\mathfrak{M}$ of $\mathfrak{K}$ by known abstract properties of dependence relations. Choose a representative system m for $\mathfrak{m}$ in k and two representative systems $M \supset m$ and $M' \supset m$ for $\mathfrak{M}$ in K and K' respectively. Denote by $\mathfrak{L}$ the minimal perfect field $\mathfrak{K}^{p^{-\infty}}$ containing $\mathfrak{K}$ and use the Teichmüller embedding process to obtain unramified complete fields $L \supset K$, $L' \supset K'$ both with the res. field $\mathfrak{L}$ and containing M and M' respectively as multiplicative representatives. Both L and L' are unramified over the $\mathfrak{p}$-adic field k_0 with perfect res. field given in the hypothesis (if k is p-adic, let k_0 be the subfield of p-adic numbers). By the equivalence theorem for perfect residue class extensions (Theorem 5) there is a map T of L on L' leaving res. classes fixed and hence carrying the multiplicative representatives M of $\mathfrak{M}$ into M'. Furthermore K^T is a subfield of L' with res. field $\mathfrak{K}$ and containing $M' = M^T \subset K^T$, so that K^T and K' have both M' and a prime element π of k_0 in common. By Lemma 4, $K^T = K'$, so that T is actually a map of K on K'. By a similar argument, $k^T = k$, and T is an automorphism of k leaving the residue classes, the prime element π of k_0 and the multiplicative representatives m of the p-basis $\mathfrak{m}$ fixed. Hence T leaves k elementwise fixed, as in Lemma 5, and T does yield the required equivalence of K to K' over k.

Though we do not have a converse proof that preservation of p-independence is a necessary condition for the equivalence of unramified extensions K and K', some restrictive condition on $\mathfrak{K}/\mathfrak{k}$ is needed to establish this equivalence, as is indicated by the following partial converse.

Theorem 9. *Let the discrete field k have a residue class field $\mathfrak{k}$ with an algebraic extension $\mathfrak{K}$. All unramified complete extensions of k with the fixed residue class field $\mathfrak{K}$ are equivalent over k if and only if $\mathfrak{K}$ is separable over $\mathfrak{k}$.*

The proof will be applicable not only to a discrete field k, but also to any field k with a value group Γ such that $\Gamma \neq p\Gamma$, where $p\Gamma$ denotes the set of all

multiples $p\gamma$ of elements γ in Γ. That separability suffices for equivalence was shown in Theorem 4. Assume conversely that $\mathfrak{R}/\mathfrak{k}$ is not separable, let $\mathfrak{k}_s$ be the maximal subfield of $\mathfrak{R}$ separable over $\mathfrak{k}$ and construct by Theorem 1 an unramified extension k_s of k with the extended res. field $\mathfrak{k}_s$. Pick an α in $\mathfrak{k}_s$ with $\alpha^{1/p}$ in $\mathfrak{R}$ but not in $\mathfrak{k}_s$, select any representative a of α in k_s, and form the algebraic extensions $L_1 = k_s(b_1)$, $L_2 = k_s(b_2)$ in which $b_1^p = a$, $b_2^p = a(1 + \pi)$. Since $\Gamma \neq p\Gamma$, π can be chosen in k so that

$$0 < V(\pi) \leqq 1, \qquad V(\pi) \text{ not in } p\Gamma. \tag{6-1}$$

The valuation of k_s has a unique extension to each field L_i, and in each case the res. field of L_i is $\mathfrak{k}_s(\alpha^{1/p})$. Unramified extensions $K_i \supset L_i$ then exist with the extended residue class field $\mathfrak{R}^* \supset \mathfrak{k}_s(\alpha^{1/p})$.

Were all unramified extensions corresponding to $\mathfrak{R}$ equivalent, there would be a map T of K_1 on K_2 leaving $\mathfrak{R}$ and k and hence (end of Theorem 3) leaving k_s elementwise fixed. The construction of b_i makes $(b_2/b_1^T)^p = 1 + \pi$ in K_2, wherefore $\pi = -1 + (1 + u)^p$, with $u = 1 - b_2/b_1^T$ in K_2. The binomial formula and the assumption $V(p) = 1$ entail

$$\pi = -1 + 1 + pu + \cdots + pu^{p-1} + u^p,$$

$$V(\pi) = \text{Min}\,[1 + V(u), \cdots, 1 + (p-1)V(u), pV(u)];$$

consequently $V(u) > 0$ and either $V(\pi) = 1 + V(u) > 1$ or else $V(\pi) = pV(u)$, both counter to (6-1). The non-uniqueness so established is the central reason for the non-uniqueness of certain maximally perfect extensions of valuations of higher rank (Mac Lane [4], §7, Example II).

7. Subfields of complete fields

When does a given p-adic field have a p-adic subfield with a specified res. field? A first answer is

THEOREM 10. *For any subfield $\mathfrak{k}$ of the residue class field $\mathfrak{R}$ of a p-adic field K there is a p-adic subfield $k \subset K$ with residue class field $\mathfrak{k}$. If $\mathfrak{k}$ is perfect, there is only one such field. More generally, if $\mathfrak{R}/\mathfrak{k}$ preserves p-independence, any two p-adic subfields k and k' with residue class field $\mathfrak{k}$ are conjugate under K.*

The statement that k and k' are *conjugate* means that there is an analytic automorphism of K leaving residue classes fixed and carrying k into k'. To prove the existence of k, construct by Theorems 1 and 2 fields $k' \subset K'$ with the res. fields $\mathfrak{k} \subset \mathfrak{R}$ and map K' on K by the structure Theorem; the map of k' is the required subfield. If k and k' are two p-adic subfields with the same perfect res. field $\mathfrak{k}$, then k and k' contain multiplicative representative systems R and R' respectively. These systems can be characterized uniquely in K by Lemma 2, hence $R = R'$. Any element in k or k' is then the limit of a series $\sum a_i p^i$ with coefficients a_i in R, so that[9] $k = k'$.

Finally, if $\mathfrak{R}/\mathfrak{k}$ preserves p-independence, the field k has a map T on k', by

[9] Examples of $k \neq k'$ with $\mathfrak{k}$ imperfect are discussed at the end of §8.

the structure Theorem, and the map may be extended to a map of $K \supset k$ on a new field $K' \supset k'$. Then K'/k' and K/k' are equivalent, by Theorem 8, under some map S leaving k' elementwise fixed. The map TS is an automorphism of K carrying k into k', so that k and k' are conjugate, as in the Theorem.

The subfield-res. field correspondence $k \to \mathfrak{k}$ has further properties related to the fact that the set S of all p-adic subfields L of K, under the ordinary inclusion relation, forms a lattice.[10]

THEOREM 11. *The p-adic subfields $k \subset K$ with perfect residue class fields form a sublattice S_p of the lattice S, while all perfect subfields $\mathfrak{k}$ of the residue class field $\mathfrak{K}$ constitute a sublattice $\mathfrak{S}_p$ of the lattice $\mathfrak{S}$ of all subfields $\mathfrak{L}$ of $\mathfrak{K}$. The correspondence $L \to \mathfrak{L}$ in which each $L \subset K$ corresponds to its residue class field $\mathfrak{L}$ is a lattice isomorphism of S_p to $\mathfrak{S}_p$, because it preserves union and intersection.*

PROOF. $\mathfrak{S}_p$ is a sublattice of $\mathfrak{S}$ (Ore [7]) because in $\mathfrak{S}$ unions and intersections of perfect fields are perfect. For the intersection this is clear; while the union $\mathfrak{k}_1 \cup \mathfrak{k}_2$ contains a maximal perfect subfield $\mathfrak{k}^*$ which then must in turn contain the perfect subfields $\mathfrak{k}_1$ and $\mathfrak{k}_2$, so $\mathfrak{k}^* = \mathfrak{k}_1 \cup \mathfrak{k}_2$ is perfect.

The res. field correspondence $L \to \mathfrak{L}$ preserves certain inclusions, for if $L \to \mathfrak{L}$ and $k \to \mathfrak{k}$, then

$$\mathfrak{k} \subset \mathfrak{L} \text{ and } \mathfrak{k} \text{ perfect imply } k \subset L. \tag{7-1}$$

For, L must contain some subfield k' with the res. field $\mathfrak{k}$ of k, and by Theorem 10, k and k' must be identical.

To prove S_p a sublattice of S we must prove that unions and intersections in S of fields in S_p lie in S_p. If $k_1 \to \mathfrak{k}_1$, $k_2 \to \mathfrak{k}_2$ with $\mathfrak{k}_1$ and $\mathfrak{k}_2$ perfect, then the unique p-adic subfield k of K with res. field $\mathfrak{k}_1 \cup \mathfrak{k}_2$ is the union of k_1 and k_2 in S. For, $k \supset k_1$ and $k \supset k_2$ by (7-1), while any p-adic subfield $k' \supset k_1 \cup k_2$ must have a res. field $\mathfrak{k}' \supset \mathfrak{k}_1 \cup \mathfrak{k}_2$, so that $k' \supset k$, whence $k = k_1 \cup k_2$. On the other hand, let $k_1 \cap k_2$ be the intersection in S of the fields k_1 and k_2, while k_0 is the unique p-adic subfield of K with the perfect res. field $\mathfrak{k}_1 \cap \mathfrak{k}_2$. By (7-1), both k_1 and k_2 contain k_0, so $k_1 \cap k_2 \supset k_0$. However, the res. field of $k_1 \cap k_2$ is contained in that of k_0, so the unramified extension $k_1 \cap k_2 \supset k_0$ is necessarily k_0, and the intersection $k_1 \cap k_2 = k_0$ has the perfect res. field $\mathfrak{k}_0 = \mathfrak{k}_1 \cap \mathfrak{k}_2$, as asserted.

Finally, the correspondence $k \to \mathfrak{k}$ is one-to-one from S_p to $\mathfrak{S}_p$ by Theorem 10, and preserves the inclusion relation by (7-1), so that it is indeed a lattice isomorphism.

Fields intermediate to given unramified extensions suggest analogous questions:

THEOREM 12. *Let $\mathfrak{p}$-adic fields $k \subset K$, K unramified over k, have residue class fields $\mathfrak{k} \subset \mathfrak{K}$ with an intermediate field $\mathfrak{L}$, $\mathfrak{k} \subset \mathfrak{L} \subset \mathfrak{K}$. A complete field L with residue class field $\mathfrak{L}$ and with $k \subset L \subset K$ will exist if $\mathfrak{L}$ has a separating trans-*

[10] In S the union $L_1 \cup L_2$ is the smallest complete field containing L_1 and L_2, and not necessarily the smallest subfield of K containing L_1 and L_2.

cendence basis over $\mathfrak{k}$ (*cf.* §3) *or if* $\mathfrak{K}/\mathfrak{k}$ *preserves p-independence and k is p-adic* (see[11] §6). *If* $\mathfrak{L}^p(\mathfrak{k}) = \mathfrak{L}$ *at most one such field L exists.*

PROOF. If $\mathfrak{L}$ has a sep. trans. basis over $\mathfrak{k}$, construct an unramified extension L' of k with res. field $\mathfrak{L}$, and map L' in K by Theorem 3; this gives the desired field L as the map of L'. If $\mathfrak{K}/\mathfrak{k}$ preserves p-independence, construct p-adic extensions $k \subset L' \subset K'$ with the res. class extensions $\mathfrak{k} \subset \mathfrak{L} \subset \mathfrak{K}$, and again map K' into K by the equivalence Theorem 8 of §6. Finally the hypothesis $\mathfrak{L}^p(\mathfrak{k}) = \mathfrak{L}$ is equivalent to the assumption that $\mathfrak{k}$ contains[12] a p-basis $\mathfrak{M}$ of $\mathfrak{L}$, which implies that k contains a representative system M of $\mathfrak{M}$. Any two intermediate fields L and L' with res. field $\mathfrak{L}$ must then contain k and M, so that Lemma 4 of §5 implies $L = L'$.

Such intermediate subfields L may not exist in cases when the hypotheses of Theorem 12 fail.

EXAMPLE I. Let $\mathfrak{K}/\mathfrak{k}$ be a finite inseparable algebraic extension with an infinite number of intermediate fields, and let k be p-adic with res. field $\mathfrak{k}$. A finite, algebraic, and unramified extension K/k with res. field $\mathfrak{K}$ can be constructed as in Theorem 1. This extension K is complete[13] and hence p-adic. The finite separable extension K/k contains only a finite number of intermediate fields L, which cannot correspond to the infinite set of intermediate fields $\mathfrak{k} \subset \mathfrak{L} \subset \mathfrak{K}$.

EXAMPLE II. Even when there are but a finite number of fields $\mathfrak{L}$ there may not exist corresponding intermediate fields L. We now construct such an extension K/k with res. fields of relative degrees $[\mathfrak{K} : \mathfrak{L}] = [\mathfrak{L} : \mathfrak{k}] = 2$, but where no intermediate p-adic field corresponding to $\mathfrak{L}$ can exist. Let $\mathfrak{P}$ be a perfect field of characteristic 2, $\mathfrak{k} = \mathfrak{P}(\tau)$ a simple transcendental extension of $\mathfrak{P}$, k a p-adic field with res. field $\mathfrak{k}$, and t a representative of τ in k. Set $K = k(\theta)$, where θ is a root of

$$f(x) = x^4 + 2x^3 + 2x^2 + t,$$

which is irreducible (mod 2) and hence irreducible over k. In the extension of the valuation V to K, θ has the residue class $\sqrt[4]{\tau}$ and K is complete and unramified with res. field $\mathfrak{K} = \mathfrak{k}(\sqrt[4]{\tau})$. Then $\mathfrak{L} = \mathfrak{k}(\sqrt{\tau})$ is an inseparable quadratic intermediate field which cannot be the res. field of an intermediate L simply because there is no quadratic field between K and k. This is the case because the resolvent cubic[14]

$$g(y) = y^3 - 2y^2 - 4ty + 4t$$

of the defining quartic equation $f(x) = 0$ is irreducible. The irreducibility of g

---

[11] If K/k is unramified and $\mathfrak{p}$-adic but not p-adic, the result still holds under the added hypothesis in the last sentence of Theorem 8.

[12] This will be the case, for instance, if $\mathfrak{L}/\mathfrak{k}$ is separable algebraic or if $\mathfrak{L}$ is obtained by adjoining to $\mathfrak{k}$ a transcendent x and all roots $x^{p^{-e}}$, etc.

[13] Cf. for instance Albert, [1], ch. XI, Theorem 11.

[14] Albert [1], ex. 9 and 17 on pp. 178–9.

may be established by the usual Eisenstein criterion,[15] for the Newton polygon of g relative to V is a straight line of slope 2/3.

This example shows that in Theorem 12 it would be impossible to state conditions on $\mathfrak{k} \subset \mathfrak{L} \subset \mathfrak{K}$ alone which would be necessary and sufficient for the existence of a corresponding intermediate p-adic field L, for it is possible to construct successively other p-adic fields $k \subset L' \subset K'$ with exactly the res. fields of this example.

II. Automorphism Groups

8. The inertial group

$G = G(K)$ denotes the group of all analytic automorphisms of a given p-adic field K and $G_1(K)$ the subgroup of "inertial" automorphisms T; that is, of those T which, as in (2-4), generate the identity automorphism on the res. field. Since K is complete in a discrete valuation it is not algebraically closed and hence not complete in any other (non-equivalent) valuation.[16] Therefore any automorphism of K is necessarily analytic, and $G(K)$ is the group of all automorphisms of K.

$G_1(K)$ is a normal subgroup of $G(K)$, exactly as for the Hilbert inertial group in an algebraic number field. We now seek the structure of the quotient group G/G_1 in the more general case when G consists only of automorphisms "relative" to a subfield k (i.e., leaving k elementwise fixed).

Theorem 13. *If $K \supset k$ are p-adic fields*[11] *with residue class fields $\mathfrak{K}/\mathfrak{k}$ preserving p-independence* (cf. §6), *and if $G(K/k)$ is the group of relative automorphisms of K over k and $G_1(K/k)$ its inertial subgroup, then the quotient group $G(K/k)/G_1(K/k)$ is isomorphic to the group $\mathfrak{G}(\mathfrak{K}/\mathfrak{k})$ of relative automorphisms of the residue extension $\mathfrak{K}$ over $\mathfrak{k}$. The isomorphism is given by the correspondence $T \to \tau$ in which τ is the residue class automorphism generated by T.*

Proof. The correspondence $T \to \tau$, where τ is given as in (2-4) by

$$\tau[H(a)] = H(a^\tau) \qquad (V(a) \geqq 0) \tag{8-1}$$

is certainly a homomorphism of G on a subgroup of $\mathfrak{G}$, and under this homomorphism the subgroup of G mapped on the identity is exactly G_1. It remains to construct a T for any given τ. If $\{V, H\}$ is the given valuation of K, then $\{V, \tau^{-1}H\}$ is another valuation of K with a different residue class homomorphism $H' = \tau^{-1}H$ but with the same res. field $\mathfrak{K}$. The two complete fields K with $\{V, H\}$ and K with $\{V, \tau^{-1}H\}$ are extensions of k with the same res. extension, so must be equivalent, by Theorem 8, under a map T of K on itself which is the identity on $\mathfrak{K}$. Hence, as in (2–5)

$$Ha = (\tau^{-1}H)(a^\tau)$$

for any a with $V(a) \geqq 0$. In other words, (8-1) holds, as desired.

[15] Mac Lane [5], or papers of Ore there mentioned.

[16] F. K. Schmidt, [9]. Our definition of equivalence of valuations and of analytic isomorphisms can be shown equivalent to those used by Schmidt; cf. Mac Lane, [4], footnote 15.

A special case, with k = the minimal complete subfield, is:

COROLLARY. *If G and G_1 are respectively the automorphism and inertial groups of a p-adic field, then G/G_1 is isomorphic to the automorphism group of the residue class field.*

Note that Theorem 13 includes also the known result that the Galois group of an unramified algebraic extension of a p-adic number field (with finite res. field) is isomorphic to the Galois group of the res. extension.

For further investigations we use the following systematic construction of automorphisms T of $G_1(K/k)$, where K is an unramified extension[11] of k with res. fields $\mathfrak{R}/\mathfrak{k}$.

LEMMA 6. *Relative inertial automorphs. If $\mathfrak{m}$ is a p-basis of $\mathfrak{k}$ contained in a p-basis $\mathfrak{m} + \mathfrak{M}$ of $\mathfrak{R}$, with $\mathfrak{m}$ and $\mathfrak{M}$ disjoint, and if M and M' are any two systems of representatives of $\mathfrak{M}$ in K, then there exists one and only one automorphism T of $G_1(K/k)$ which carries M into M'. All automorphisms T of $G_1(K/k)$ can be obtained thus, for fixed M and variable M'.*

If K is p-adic, this Lemma also gives the whole group $G(K)$.

PROOF. By Theorem 6 embed K in complete unramified extensions L and L' with res. field $\mathfrak{L} = \mathfrak{R}(\mathfrak{M}^{p^{-\infty}})$ so that M and M' respectively are multiplicative representatives of $\mathfrak{M}$. The res. field $\mathfrak{L}$ has as p-basis the p-basis $\mathfrak{m}$ of $\mathfrak{k}$, by Lemma 3 of §5, so L/k preserves p-independence and the fundamental equivalence Theorem 8 applies. It provides a map T of L on L' which leaves elements of $\mathfrak{L}$ and k fixed and which carries the multiplicative representatives M into M'. Pick a system m of representatives of $\mathfrak{m}$ in k. The subfield K^T of L' contains k and M', hence contains the representative system $m + M'$ of the p-basis $\mathfrak{m} + \mathfrak{M}$ of $\mathfrak{R}$, while the subfield K originally embedded in L' contains the same set $m + M'$. Any prime element π of k is contained in both K and K^T, hence $K^T = K$ by Lemma 4 and T is an automorphism of K. By construction, T leaves all residue classes fixed, hence T is in $G_1(K/k)$.

Any relative inertial automorphism T manifestly maps M on some other representative system M', and M' uniquely determines T by Lemma 5, as asserted. A special case of this uniqueness is

THEOREM 14. *A p-adic field with a perfect residue class field $\mathfrak{R}$ has except for the identity no automorphism which leaves every residue class fixed.*

The lemma may also be applied to show that the subfield $k \subset K$ with specified res. field $\mathfrak{k}$ constructed in Theorem 10 need not be unique in all cases when $\mathfrak{k}$ is not perfect. Specifically, if the p-adic field K has a res. field $\mathfrak{R}$ with a proper subfield $\mathfrak{k}$ containing some element ξ not in $\mathfrak{R}^p$, then K contains distinct p-adic subfields with res. field $\mathfrak{k}$. For there is a p-basis $\mathfrak{M}$ of $\mathfrak{R}$ containing the element ξ not in $\mathfrak{R}^p$, by known properties of p-independence. Let $k \subset K$ be a p-adic subfield with res. field $\mathfrak{k}$ and let M be any system of representatives for $\mathfrak{M}$ in K containing a representative x of ξ. Since $k < K$ there is a b in K but not in k with $Vb > 0$, so that $y = x + b$ is another representative of ξ in K. The replacement of x in M by this new representative y yields a new system M' of representatives of $\mathfrak{M}$. By Lemma 6 there is a map T of $G_1(K)$ with $M^T =$

M'. Under T the map k^T of the subfield k is another p-adic subfield with res. field $\mathfrak{k}$, and $k^T \neq k$ because k^T contains the element y in M' but not in k.

9. The series of ramification groups

If K/k is unramified and p-adic, denote by $G_n(K/k)$ the group of all those relative automorphisms T of K/k for which b in K and $V(b) \geqq 0$ imply $b^T \equiv b \pmod{\pi^n}$. For $n = 1$, G_n is the inertial group already considered. If $n > 1$, $G_n(K/k)$ is a *pseudo-ramification group* because its formal definition is like that for the usual Hilbert ramification group, although it cannot correspond to any (non-existent!) ramification of K over k. As in the classical case, one proves that each $G_n(K/k)$ is a normal subgroup of $G_1(K/k)$. The successive quotient groups are abelian, as for the ordinary ramification groups, and have the following explicit structure:

THEOREM 15. *Let K/k be p-adic[11] with residue class fields $\mathfrak{K}/\mathfrak{k}$ preserving p-independence and with pseudo-ramification groups $G_n = G_n(K/k)$, $n \geqq 1$. Then each quotient group G_n/G_{n+1} is abelian and is isomorphic to the direct product $\prod \mathfrak{K}^+$ of $\mathfrak{c}$ abelian groups, where each factor $\mathfrak{K}^+$ is isomorphic to the additive group of the residue field $\mathfrak{K}$, while the cardinal number $\mathfrak{c}$ of factors is the cardinal number of any set[17] $\mathfrak{N}$ which has the property that there is a p-basis $\mathfrak{m}$ of $\mathfrak{k}$ such that $\mathfrak{N}$ and $\mathfrak{m}$ are disjoint and $\mathfrak{N} + \mathfrak{m}$ is a p-basis of $\mathfrak{K}$.*

COROLLARY. *A p-adic field K with pseudo-ramification groups G_n has each G_n/G_{n+1} abelian and isomorphic to the direct product of $\mathfrak{c}$ factors $\mathfrak{K}^+$, where $\mathfrak{c}$ is the degree of imperfection of the residue class field $\mathfrak{K}$.*

The *degree of imperfection* of $\mathfrak{K}$ is by definition the cardinal number of elements in a p-basis of $\mathfrak{K}$; it is independent of the choice of the p-basis, as shown by Teichmüller [11].

PROOF OF THEOREM 15. For each residue class λ of the set $\mathfrak{N}$ pick a representative d_λ in K. By definition, G_n consists of maps T for which

$$d_\lambda^T = d_\lambda + \pi^n t_\lambda \tag{9-1}$$

for some elements t_λ in the valuation ring of K. The correspondence $T \to \{H(t_\lambda)\}$ carries G_n into the ordered set $\{H(t_\lambda)\}$ of $\mathfrak{c}$ residue classes $\alpha_\lambda = H(t_\lambda)$, where λ ranges over all $\mathfrak{c}$ elements of $\mathfrak{N}$. If $\mathfrak{K}_\lambda^+$ is the additive group of all residues α_λ, this correspondence is a homomorphism $G_n(K/k) \sim \mathfrak{K}^+$ because $S \to \{H(s_\lambda)\}$, $T \to \{H(t_\lambda)\}$ yields by definition (9-1) $ST \to \{H(t_\lambda + s_\lambda^T)\} = \{H(t_\lambda + s_\lambda)\}$. The correspondence carries T's in G_{n+1}, and only these (see Lemma 7, below) into the identity. Finally, every possible element $\{\alpha_\lambda\}$ of the direct product is the correspondent of some T, for by Lemma 6 there is an automorphism T with $d_\lambda^T = d_\lambda + \pi^n a_\lambda$, where a_λ is any representative of α_λ. That this T must lie in G_n is a result of the following Lemma, which can be readily established by a decomposition proof resembling that of Lemma 5 in §5:

LEMMA 7. *In a p-adic[11] field $K \supset k$, an inertial automorphism T of $G_1(K/k)$*

[17] There always is such a set $\mathfrak{N}$, because $\mathfrak{K}/\mathfrak{k}$ preserves p-independence. It can be shown that the cardinal number $\mathfrak{c}$ of $\mathfrak{N}$ is independent of the choice of $\mathfrak{N}$ in $\mathfrak{K}/\mathfrak{k}$.

is in the pseudo-ramification group $G_n(K/k)$ if and only if $d^T \equiv d \pmod{p^n}$ holds for all d of a fixed system M of representatives of a p-basis $\mathfrak{M}$ of $\mathfrak{K}$.

10. The inertial subfield

Each ramification group G_n of a p-adic field determines a corresponding subfield consisting of all elements of the field invariant under G_n. Unlike the classical case, these pseudo-ramification fields all coincide.

THEOREM 16. *In a p-adic field K let k be the unique p-adic subfield (cf. Theorem 10 in §7) whose residue class field $\mathfrak{k}$ is the maximal perfect subfield of the residue class field $\mathfrak{K}$ of K, and let G_m be the m^{th} pseudo-ramification group of K, for $m \geqq 1$. Then k consists of all those and only those elements of K invariant under all the automorphisms of G_m.*

That any T in G_m leaves k elementwise fixed is immediate, for the unique subfield k can only go into itself under T, so that T must leave the elements of k fixed as in Theorem 14. The crux of the theorem is the converse statement, that for any element b in $K - k$ there is a T of G_m with $b^T \neq b$. This will be established by explicit construction of a T as in Lemma 6. After some reductions, we can assume that the given b in the valuation $\{V, H\}$ of K has $V(b) = 0$ and $H(b) = \beta$ for some β not in $\mathfrak{k}$. Hence the query: given b, is there a T in G_m such that

$$Vb = 0,\ Hb = \beta \text{ in } K - k_0 \text{ imply } b^T \neq b? \tag{10-1}$$

First, a p-basis on which to build T! Since β is not in $\mathfrak{k}$, the maximal perfect subfield obtainable as the intersection of all fields[18] $\mathfrak{K}^{p^n}$, there is some index u for which β is in $\mathfrak{K}^{p^u}$, but not in $\mathfrak{K}^{p^{u+1}}$. For some γ,

$$\beta = \gamma^{p^u}, \qquad \gamma \text{ in } \mathfrak{K} \text{ but not in } \mathfrak{K}^p. \tag{10-2}$$

There is then a p-basis $\{\gamma\} + \mathfrak{M}$ of $\mathfrak{K}$ containing this element γ. Fix on a representative c of γ in K and a representative system M for $\mathfrak{M}$ in K, and construct by Lemma 6 of §8 the relative automorphism T of K/k with

$$c^T = c(1 + p^m), \qquad d^T = d \qquad (\text{all } d \text{ in } M). \tag{10-3}$$

By Lemma 7 this T is in $G_m(K)$. We propose to show T effective in (10-1).

T leaves fixed the representatives of M and hence all elements of the subfield L which is constructed as the complete closure in K of the field $k(M)$. The res. field $\mathfrak{L}$ of L is $\mathfrak{L} = \mathfrak{k}(\mathfrak{M})$. For, the elements of the p-independent set $\mathfrak{M}$ are algebraically independent over $\mathfrak{k}$ (Teichmüller [11]), so that the representing elements of M are necessarily algebraically independent over k. Furthermore the residue-class homomorphism H of $k(M)$ on $\mathfrak{L}$ is given by a formula like (4-1) and (4-3), so $\mathfrak{L} = \mathfrak{k}(\mathfrak{M})$, as asserted.

$\mathfrak{L}$ has certain extensions $\mathfrak{K}_n = \mathfrak{K}^{p^n}(\mathfrak{M})$ such that (Teichmüller [10], Lemma 11)

$$\mathfrak{K}_n = \mathfrak{L}(\mathfrak{K}^{p^n}), \qquad \mathfrak{K} = \mathfrak{K}_n(\gamma), \qquad [\mathfrak{K} : \mathfrak{K}_n] = p^n. \tag{10-4}$$

---

[18] Here $\mathfrak{K}^{p^n}$ denotes the field of all p^n powers ξ^{p^n}, for ξ in $\mathfrak{K}$.

Thus γ functions as a "relative p-basis" of $\mathfrak{K}/\mathfrak{L}$. We obtain representatives of all residue classes of $\mathfrak{K}_n$ in terms of the valuation rings B and B_L of K and L, respectively and of the set B^{p^n} of all p^{nth} powers of elements of B, as follows.

LEMMA 8. *Any element a of the ring $R_n = B_L[B^{p^n}]$ has $a^T \equiv a \pmod{p^{n+m}}$ and $H(a)$ in $\mathfrak{K}_n$. Conversely, any element of $\mathfrak{K}_n$ has a representative in R_n.*

PROOF. By the definition (10-4), $\mathfrak{K}_n = \mathfrak{K}^{p^n}(\mathfrak{L}) = \mathfrak{K}^{p^n}[\mathfrak{L}]$ has a set of representatives in the correspondingly constructed ring $R_n = B_L[B^{p^n}]$. On the other hand, any a is a polynomial $f(B^{p^n})$ with coefficients in B_L, hence $H(a)$ is in $\mathfrak{K}_n$ and $a^T \equiv a \pmod{p^{n+m}}$ by Lemma 1.

The original element b will subsequently be represented by polynomials in the quantity c of (10-3) with coefficients in R_n. Hence the following computation of the effect of T on such a polynomial:

LEMMA 9. *If the polynomial $F(y)$ is in $R_n[y]$, where $n \geqq i + 1 \geqq 1$, then*

$$[F(c^{p^i})]^T - F(c^{p^i}) \equiv c^{p^i} p^{i+m} F'(c^{p^i}) \pmod{p^{i+m+1}}, \tag{10-5}$$

where $F'(y)$ is the derivative of $F(y)$. In particular, if $i \geqq 0$,

$$(c^{p^i})^T - c^{p^i} \equiv c^{p^i} p^{i+m} \pmod{p^{i+m+1}}. \tag{10-6}$$

PROOF. The relation (10-6) results from the construction $c^T = c(1 + p^m)$ of T in (10-3) and from the binomial expansion

$$(1 + p^m)^{p^i} \equiv 1 + p^{m+i} \pmod{p^{m+i+1}}.$$

This congruence, together with Lemma 9, fails when $p = 2$ and $m = 1$, but this exception is avoided by assuming $m \geqq 2$; for the construction of a T in G_2 certainly provides a T in the larger group G_1! The congruence (10-5) is derived from the Taylor expansion of $[F(c^{p^i})]^T \equiv F(c^{p^i} + p^{i+m} c^{p^i})$, using (10-6) and Lemma 8 to find the effect of T on coefficients of F.

The elements a left fixed by T will next be sought. Suppose a in B and $a^T \equiv a \pmod{p^{m+s}}$ for some integer $s \geqq 1$. The res. $H(a) = \alpha$ is by (10-4) in $\mathfrak{K}_s(\gamma)$, so there is a polynomial $g(y)$ in $\mathfrak{K}_s[y]$ of degree less than p^s with $\alpha = g(\gamma)$. Then $a \equiv G(c) \pmod{p}$, where $G(y)$ is a polynomial whose coefficients are in R_s and represent corresponding coefficients of $g(y)$, as in Lemma 8. Therefore $a - G(c) = pa_1$, for some a_1 in B, and

$$0 \equiv a^T - a \equiv [G(c)]^T - G(c) + p(a_1^T - a_1) \pmod{p^{m+1}}.$$

But $a_1^T \equiv a_1 \pmod{p^m}$ because T is in G_m, so (10-5) for $i = 0$ gives

$$0 \equiv a^T - a \equiv cp^m G'(c) \pmod{p^{m+1}}.$$

Consequently $cG'(c) \equiv 0 \pmod{p}$ and $\gamma g'(\gamma) = 0$. This gives an equation $yg'(y) = 0$ of degree less than p^s for γ over $\mathfrak{K}_s$, which contradicts (10-4) unless $g'(y) = 0$ identically. In this event, g has the form $g(y) = f(y^p)$ and as a result $G(y) = F(y^p)$, provided G was originally so chosen that each zero coefficient of g corresponds to a zero coefficient in G. Thus $a \equiv F(c^p) \pmod{p}$, where F has degree less than p^{s-1} and coefficients in R_s. Repeating the process for higher moduli, we finally obtain

LEMMA 10. *If the element a in B has $a^T \equiv a \pmod{p^{m+s}}$, and if $0 < n \leqq s$, then a has an expansion*

$$a \equiv F_n(c^{p^n}) + F_{n-1}(c^{p^{n-1}})p + \cdots + F_1(c^p)p^{n-1} \pmod{p^n}, \tag{10-7}$$

where each $F_i(y)$ is a polynomial in $R_{s-n+i}[y]$ of degree less than p^{s-n}, and where any coefficient of F_i congruent to 0 (mod p) is equal to 0.

The proof is by induction on n for fixed s. The case $n = 1$ has just been established. Assume next that (10-7) holds for some n and that $s \geqq n + 1$. The next term in this expansion could be written as $a_n p^n$, for some a_n in B. We can again write $a_n \equiv F_0(c)$ where $F_0(y)$ is in $R_{s-n}[y]$ and of degree less than p^{s-n}. The expansion now is

$$a = F_n(c^{p^n}) + \cdots + F_1(c^p)p^{n-1} + F_0(c)p^n + a_{n+1}p^{n+1}, \tag{10-8}$$

for some element a_{n+1} in B, so that $a_{n+1}^T \equiv a_{n+1} \pmod{p^m}$. The difference $a^T - a$ is $\equiv 0 \pmod{p^{m+n+1}}$ by hypothesis and may be calculated by applying Lemma 9 to (10-8). The congruence $\sum c^{p^i} F_i'(c^{p^i}) \equiv 0 \pmod{p}$ results, where the sum runs from $i = 0$ to n. The homomorphic image of this congruence becomes

$$\gamma^{p^n} f_n'(\gamma^{p^n}) + \cdots + \gamma^p f_1'(\gamma^p) + \gamma f_0'(\gamma) = 0, \tag{10-9}$$

where $\gamma = H(c)$ and $f_k(y)$ denotes the H-image of $F_k(y)$.

An induction on j from this equation shows that $f_j'(y) = 0$. The initial case $j = 1$ can be subsumed in the induction by starting with the general assumption that the polynomial $h_j(z)$ determined by

$$z^{p^j} h_j(z) = z^{p^n} f_n'(z^{p^n}) + \cdots + z^{p^j} f_j'(z^{p^j}) \tag{10-10}$$

has a root $z = \gamma$; for when $j = 0$ this assumption is (10-9). Each $f_i(z)$, as image of $F_i(z)$, has coefficients in $\mathfrak{R}_{s-n+i}$, so that $h_j(z)$ has coefficients in $\mathfrak{R}_r$, where $r = s - n + j > 0$. Over this field $\mathfrak{R}_r$, γ must by (10-4) satisfy the irreducible equation $z^{p^r} - \gamma^{p^r} = 0$, for which reason the other equation $h_j(z) = 0$ for γ is a multiple

$$h_j(z) = q(z)(z^{p^r} - \gamma^{p^r}). \tag{10-11}$$

In (10-11) the terms on the left are of two sorts: first, the terms $z^{p^i - p^j} f_i(z^{p^i})$ from (10-10) with $i > j$, which are sums of terms of the form dz^e of degree $e = p^i - p^j + e_0 p^i \equiv -p^j \pmod{p^{j+1}}$; secondly, terms from $f_j'(z^{p^j})$ in which no terms z^e with $e \equiv -p^j \pmod{p^{j+1}}$ can occur, because f_j' is a derivative in a field of characteristic p. The terms of this second type have a degree $e' = e_0' p^j < p^{s-n} p^j = p^r$, less than the minimum degree of terms in $q(z)z^{p^r}$. One argues from this circumstance that $q(z)$ can contain only terms z^ν with exponents $\nu \equiv -p^j \pmod{p^{j+1}}$ and therefore that $f_j'(z^{p^j})$, which is in (10-11) the only source of terms not satisfying this congruence condition, must be identically zero. This conclusion, $f_j'(y) = 0$, completes the induction.

This identity $f_i'(y) = 0$ implies that $f_i(y) = g_{i+1}(y^p)$ for some polynomial $g_i(z)$ in $\mathfrak{R}_{s-n+i}[z]$ of degree less than p^{s-n-1}. The hypothesis concerning zero

coefficients for the F's, inclusing of a proper such choice for the F_0 constructed at the start of the proof, thereby insures that $F_i(y) = G_{i+1}(y^p)$, for G_{i+1} some polynomial with coefficients in R_{s-n+i}. With these G's, (10-8) can be rewritten as the desired expansion (10-7) of a (mod p^{n+1}), so that the induction on n for the Lemma is proven.

Return to the query (10-1) and suppose that the element b there had $b^T = b$. Apply Lemma 10 to b with $n = s = u + 1$, for u chosen as in (10-2). The Lemma states that $b \equiv F_n(c^{p^n})$ (mod p), where F_n is of degree less than $p^0 = 1$, so that b must be in $R_n = R_{u+1}$ with a residue class $\beta = \gamma^{p^u}$ in $\mathfrak{R}_n = \mathfrak{R}_{u+1}$, contrary to (10-4). The second half of Theorem 16 is thus completed.

HARVARD UNIVERSITY.

BIBLIOGRAPHY

(1) A. A. ALBERT, *Modern Higher Algebra*, Chicago, 1937.

(2) H. HASSE AND F. K. SCHMIDT, *Die Struktur diskret bewerteter Körper*, Journal für die Mathematik, 170 (1934), pp. 4–63.

(3) S. MACLANE, *A Construction for Absolute Values in Polynomial Rings*, Trans. Am. Math. Soc., 40 (1936), pp. 363–395.

(4) S. MACLANE, *The Uniqueness of the Power Series Representation of certain Fields with Valuations*, Annals of Math., 39 (1938), pp. 370–382.

(5) S. MACLANE, *The Schönemann-Eisenstein Irreducibility Criteria in Terms of Prime Ideals*, Trans. Am. Math. Soc., 43 (1938), pp. 226–239.

(6) T. NAKAYAMA, *Divisionsalgebren über diskret bewerteten perfekten Körpern*, J. f. Math., 178 (1937), pp. 11–13.

(7) O. ORE, *On the Foundations of Abstract Algebra* I, Annals of Math., 36 (1935), pp. 406–437.

(8) A. OSTROWSKI, *Untersuchungen zur arithmetischen Theorie der Körper*, Math. Zeit., 39 (1934), pp. 269–404.

(9) F. K. SCHMIDT, *Mehrfach perfekte Körper*, Math. Ann., 108 (1933), pp. 1–25.

(10) O. TEICHMÜLLER, *Diskret bewertete perfekte Körper mit unvollkommenem Restklassenkörper*, J. f. Math., 176 (1937), pp. 141–152.

(11) O. TEICHMÜLLER, p-*Algebren*, *Deutsche Mathematik*, 1 (1936), pp. 362–388.

(11a) O. TEICHMÜLLER, *Über die Struktur diskret bewerteter Körper*, Göttinger Nachrichten, 1936, Math. Phys. Klasse, Fachgruppe I, vol. 1, Nr. 10.

(12) E. WITT, *Schiefkörper über diskret bewerteten Körpern*, J. f. Math., 176 (1936), pp. 153–156.

(13) E. WITT, *Zylische Körper und Algebren der Charakteristik* p *von Grade* p^n, J. f. Math., 176 (1936), pp. 126–140.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

THE EXISTENCE OF MINIMAL SURFACES OF GENERAL CRITICAL TYPES

BY MARSTON MORSE AND C. TOMPKINS

(Received February 7, 1939)

1. Introduction

We are concerned with the totality of minimal surfaces bounded by a simple closed rectifiable curve g in a euclidean space of dimension $n > 1$. We are not concerned with minimal surfaces of minimum type except in so far as they appear as special cases of the general theory, corresponding to 0-cycles in our function space. We shall show that such a theory can be developed as a particular application of the general critical point theory. The function will be the Douglas-Dirichlet [1] functional, and the independent variable an admissible transformation of the parameter of g. That such an application of the general critical point theory to the functionals of variational theory is possible has already been shown [7] for simple integrals of the positive definite type. Having in mind the possibility of an application of the critical point theory to multiple integrals Morse has recently developed the underlying abstract principles in the reference [7] cited below. This development will presently be extended from the same point of view in references [7] and [8].

We begin with a summary of the underlying theory defining the conditions of *bounded compactness*, *regularity* at *infinity*, and *weak reducibility* of F at a point p. These conditions are the simplest conditions under which the general theory can be applied. Our study of minimal surfaces is fundamentally the establishing of these conditions. We begin with the assumption that the number of minimal surfaces bounded by g is finite. In the last section this condition on the number of minimal surfaces is removed.

We give an example in which the curve g bounds two minimal surfaces of minimum type belonging to disconnected critical sets. The general theory then affirms the existence of at least one minimal surface of non-minimum type.

2. The general critical point theory

We shall here present those concepts and theorems of the general theory which are essential to the theory of minimal surfaces in the large. For proofs and further details the reader is referred to references [7] and [8].

We start with an abstract metric space M with points p, q, r etc. Distances pr, pq etc. shall satisfy the usual axioms

$$pp = 0, \qquad pq \neq 0 \quad \text{if } p \neq q, \; pq \leqq pr + qr.$$

We shall make use of Vietoris chains and cycles on M with coefficients in an

arbitrary field. The reader who desires special simplicity can take this field as the field of integers mod 2.

Let $F(p)$ be a real single-valued function of the point p on M. We suppose that F is finite and positive but in general not bounded.

Let E be a subset of M. We shall admit deformations Δ of points initially on E. The deformation Δ shall replace a point p on E at the time $t = 0$ by a point $q = q(p, t)$ on M at the time t, where p is on E and t varies on a closed interval $(0, \tau)$. We shall suppose that τ is a positive constant and that $q(p, t)$ varies continuously with its arguments. The curve $q = q(p, t)$ obtained by holding p fast and varying t will be termed the *trajectory* T defined by p. If a point q precedes a point r on T, q will be termed an *antecedent* of r.

We shall say that Δ admits a *displacement function* on E, if whenever q is an antecedent of r such that $qr > e > 0$, then

$$F(q) - F(r) > \delta(e),$$

where $\delta(e)$ is a positive, single-valued function of e. A deformation of E which possesses a displacement function on each compact subset of E will be termed an *F-deformation* of E.

If c is a finite constant the subset of points of M on which $F < c$ will be said to be *below* c, while the subset of points on which $F \leqq c$ will be said to be below $c+$.

We shall now enumerate three conditions on F and M under which the theory proceeds most simply.

I. *Bounded compactness.* Under this condition the subsets $F \leqq c$ are compact for each finite constant c.

Under the condition of bounded compactness F is lower semi-continuous. To see this let p_n be a sequence of points tending to a point q as n tends to ∞. We must show that

$$\liminf_{n \to \infty} F(p_n) \geqq F(q). \tag{2.0}$$

To that end let the left member of (2.0) have the value c. If c is infinite (2.0) holds. If c is finite then for any positive constant e and for an infinite sequence of increasing positive integers n_i, p_{n_i} lies below $c + e$. By virtue of the condition of bounded compactness, q lies below $(c + e)+$. Hence

$$F(q) \leqq c + e$$

for each positive e, and (2.0) follows.

II. *Regularity at infinity.* Corresponding to each compact subset A of M we here require the existence of a continuous deformation of A such that the final image of A is below some finite constant, and such that any subset of A on which F is bounded is deformed through a set of points on which F is bounded.

If F is regular at infinity each cycle on M is homologous to a cycle below some finite constant. A cycle z below a finite constant c and bounding on M, is bounding on a set on which F is bounded.

III. *Weak upper-reducibility.* The function F will be said to be weakly upper-reducible at a point p, if corresponding to each constant $c > F(p)$ there exists a neighborhood N_c of p relative to $F \leqq c$, and a deformation Δ of N_c, with the following properties:

(i) *For some positive η the points of N_c not initially below $c - \eta$ are deformed onto the set below $c - \eta$,* with Δ admitting a displacement function.

(ii) *The points of N_c initially below $c - \eta$ remain below $c - \eta$ throughout Δ.*

To obtain the consequences of the conditions I, II, and III we introduce the notion of a homotopic critical point and its type numbers. A point p is said to be *homotopically ordinary* if some neighborhood of p relative to the set $F \leqq F(p)$ admits an F-deformation which displaces p. A point which is not homotopically ordinary is termed *homotopically critical.*

A point set A will be said to be *definitely below a* (written d-below a) if A lies below $a - e$ for some positive e. The phrase d-mod $F < a$ shall be understood to mean mod some compact set d-below a. Let u be a k-cycle below $a+$ taken d-mod $F < a$. A homology

$$u \sim 0 \qquad (\text{on } F \leqq a)(\text{d-mod } F < a)$$

will be termed an *a-homology.* A k-cycle u below $a+$, d-mod $F < a$, not a-homologous to 0, will be termed a *k-cap* with *cap-limit a.*

A group of k-cycles or k-caps etc. will be said to have a property A if every element of the group, with the possible exception of the null element, has the property A. The group will be termed a *maximal* group with the property A if it is a proper subgroup of no group with property A. With this understood we turn to certain important definitions.

Let σ be a homotopic critical point of F with $F(\sigma) = c$. Suppose first that σ is isolated among homotopic critical points below $c+$, and let U be a neighborhood of σ relative to $F \leqq c$ at a positive distance from other homotopic critical points of F below $c+$. The dimension of a maximal group of k-caps on U with cap-limit c will be called the *k^{th} cap type number* of σ. This type number is independent of the choice of U restricted as above.

The simplest general theorem is the following:

THEOREM 2.1. *Let F be a function satisfying conditions* I, II, *and* III *and possessing at most a finite number of homotopic critical points. Let R_k be the k^{th} connectivity of M and M_k the sum of the k^{th} cap type numbers of the homotopic critical points of M. If the numbers R_k and M_k are finite and if we set $e_k = M_k - R_k$ the following relations hold*

$$\begin{aligned} e_0 &\geqq 0, \\ e_1 - e_0 &\geqq 0, \\ e_2 - e_1 + e_0 &\geqq 0, \\ e_n - e_{n-1} + \cdots + (-1)^n e_0 &= 0, \end{aligned} \tag{2.1}$$

where n is the maximum of the indices of the numbers M_k.

Any condition implying that the numbers R_k are finite is not restrictive for our purposes, inasmuch as we shall show that in our application $R_0 = 1$ and $R_k = 0$, $k > 1$. The restriction that the number of homotopic critical points be finite can be removed, as we shall see in §7. We state the following simple corollary of the theorem.

COROLLARY. *If the conditions of the theorem are satisfied, if* $R_0 = 1$, *and if there are* m *isolated proper relative minima there are at least* $m - 1$ *homotopic critical points of type* 1, *counting critical points with their type numbers.*

Under the conditions of the corollary

$$M_0 \geqq m, \qquad e_0 \geqq m - 1, \qquad e_1 = M_1 - R_1 .$$

From the relation $e_1 \geqq e_0$ we infer that

$$M_1 \geqq R_1 + m - 1,$$

and the proof is complete.

In §7 we shall point out that when conditions I, II, and III are satisfied and $R_0 = 1$, and the distribution of critical points is otherwise completely arbitrary, the existence of two disconnected minimizing critical sets implies the existence of at least one homotopic critical set σ of non-minimizing type. More precisely the first type number of σ as distinguished from the zero$^{\text{th}}$ type number will not be null. This result is a very special application of the general relations and existence theorems.

3. The Douglas functional $A(\varphi)$

We suppose that the curve g is simple, and representable in the form

$$x_i = g_i(s) \qquad (i = 1, \cdots, n) \tag{3.1}$$

in the space of the n coördinates, with s the arc length along g. We shall also assume that the derivatives $g_i'(s)$ exist and satisfy a Lipschitz condition of the form

$$| g_i'(s + \Delta s) - g_i'(s) | \leqq K | \Delta s |, \qquad (i = 1, \cdots, n) \tag{3.2}$$

where K is a constant independent of i. Without loss of generality we can suppose that the length of g is 2π, and that

$$g_i(s + 2\pi) \equiv g_i(s) \qquad (i = 1, \cdots, n).$$

We shall admit other representations of g, in particular representations of the form

$$x_i = p_i(\alpha) = g_i[\varphi(\alpha)] \tag{3.3}$$

where $\varphi(\alpha)$ is a continuous non-decreasing function of α such that

$$\varphi(\alpha + 2\pi) \equiv \varphi(\alpha) + 2\pi.$$

A function $\varphi(\alpha)$ of this type can be constant on certain intervals of the α-axis,

but the length of such intervals can never be as great as 2π. A representation of g of the form (3.3) will be termed a *p-curve* defined by φ.

Let (r, θ) be polar coördinates in a plane of rectangular coördinates (u, v). Corresponding to the curve $[p(\alpha)]$ there exists a "harmonic surface" S of the form

$$x_i = x_i(r, \theta) = \frac{1}{2\pi}\int_0^{2\pi} p_i(\alpha)\,\frac{(1-r^2)}{1-2r\cos(\theta-\alpha)+r^2}\,d\alpha. \tag{3.4}$$

The functions $x_i(r, \theta)$ are harmonic for $r < 1$ and can be so defined for $r = 1$ as to be continuous on the disc $r \leqq 1$. So extended,

$$x_i(1, \theta) \equiv p_i(\theta).$$

Let $y_i(u, v)$ be a representation of S such that

$$y_i(r\cos\theta, r\sin\theta) = x_i(r, \theta).$$

Let H represent the region $u^2 + v^2 < 1$. We suppose that the Dirichlet integral

$$D(p) = \tfrac{1}{2}\int\int_H \sum_i \left[\left(\frac{\partial y_i}{\partial u}\right)^2 + \left(\frac{\partial y_i}{\partial v}\right)^2\right] du\,dv$$

is finite. We term $D(p)$ the Dirichlet integral *defined* by $[p(\alpha)]$ or by $\varphi(\alpha)$. According to Douglas $D(p)$ equals the improper integral

$$A(\varphi) = \frac{1}{16\pi}\int\int_Q \sum_i \left|\frac{p_i(\alpha) - p_i(\beta)}{\sin\dfrac{(\alpha-\beta)}{2}}\right|^2 d\alpha\,d\beta$$

extended over the square Q

$$0 \leqq \alpha \leqq 2\pi, \qquad 0 \leqq \beta \leqq 2\pi.$$

Let the region $u^2 + v^2 < 1$ be subjected to a $1-1$ conformal transformation T into itself. In this paper the term conformal shall mean either directly or inversely conformal. Under T the circle $r = 1$ undergoes a homeomorphism. A point with angular coördinate α is thereby transformed into a point with angular coördinate $\theta(\alpha)$. We set

$$p_i[\theta(\alpha)] = q_i(\alpha)$$

and recall the well known fact that

$$D(p) = D(q).$$

Recall also that there exist transformations of the type T in which three arbitrary distinct points of the circle $r = 1$ are carried into three arbitrary distinct points of $r = 1$. These facts introduce and partially justify the three point condition on the transformations $\varphi(\alpha)$.

Under the three point condition on $\varphi(\alpha)$ the relation $\alpha = \varphi(\alpha)$ shall hold for three given distinct values α_1, α_2, α_3 of α on the interval $0 \leqq \alpha < 2\pi$.

The values α_1, α_2, α_3 are held fast throughout the remainder of this paper. When φ satisfies the three point condition the harmonic surface S defined by φ is represented as a continuous image of the disc $r \leqq 1$ in such a manner that the three points on the boundary of the disc at which θ equals α_1, α_2, and α_3 correspond respectively to the three points of g at which s equals α_1, α_2, and α_3.

The space Ω. The space M of §2 is here taken as a space Ω defined as follows: Ω shall consist of those transformations $\varphi(\alpha)$ which satisfy the three point condition and for which $A(\varphi)$ is finite. It is understood that φ is continuous and non-decreasing as previously. We term φ a point of Ω. To two points φ and ψ of Ω we assign the distance

$$\varphi\psi = \max. \mid \varphi(\alpha) - \psi(\alpha) \mid \qquad (0 \leqq \alpha \leqq 2\pi).$$

The space Ω so defined possesses a metric satisfying the usual axioms. The function $A(\varphi)$ now replaces the function F of the general theory. The function $A(\varphi)$ is finite at each point of Ω. We turn next to the conditions I, II, and III on $A(\varphi)$. The condition of bounded compactness here takes the form

The subset of points φ *of* Ω *for which* $A(\varphi)$ *is at most a finite constant is compact.*

That this condition is satisfied follows from the work of Douglas. See also [5] and [3] p. 78. As pointed out in §2 this condition implies that $A(\varphi)$ is lower semi-continuous.

We supplement the preceding by a definition of a distance between two p-curves $[p(\alpha)]$ and $[q(\alpha)]$ representing g. We take this distance as the maximum of the distance

$$\left\{\sum_i [p_i(\alpha) - q_i(\alpha)]^2\right\}^{\frac{1}{2}}$$

as α varies from 0 to 2π inclusive. The space of p-curves representing g is thus a metric space of the usual type.

We now turn to condition II of §2.

4. Regularity at infinity

The integral $A(\varphi)$ was taken over the square Q in the (α, β)-plane. On Q the integral is singular on the diagonal $\alpha = \beta$ as well as at the vertices of Q. We shall find it simpler to replace Q by the parallelogram ω on which

$$0 \leqq \beta \leqq 2\pi,$$

$$\beta - \pi \leqq \alpha \leqq \beta + \pi.$$

On ω the integrand of $A(\varphi)$ is singular only on the line $\alpha = \beta$.

Let $\theta(\alpha)$ satisfy all the conditions on a point $\varphi(\alpha)$ of Ω except the condition that $A(\varphi)$ be finite. We introduce the integral

$$H(\theta) = \frac{1}{16\pi} \int\int_\omega \left| \frac{\theta(\alpha) - \theta(\beta)}{\sin \dfrac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta,$$

noting that this integral is improper and understanding that it may be infinite. We shall prove the following theorem.

THEOREM 4.1. *A necessary and sufficient condition that $A(\theta)$ be finite is that $H(\theta)$ be finite.*

For a rectifiable curve the length of a chord is at most the length of an arc subtended by the chord. Hence

$$\sum_i \{g_i[\theta(\alpha)] - g_i[\theta(\beta)]\}^2 \leqq [\theta(\alpha) - \theta(\beta)]^2. \tag{4.0}$$

It follows that

$$A(\theta) \leqq H(\theta). \tag{4.1}$$

Hence the condition that $H(\theta)$ be finite is sufficient that $A(\theta)$ be finite.

To prove the condition necessary we assume that $A(\theta)$ is finite. Let e be a positive constant less than π and let ω_e be the subset of ω on which $|\alpha - \beta| < e$. We set

$$\omega = \omega_e + \omega_e^*,$$

where ω_e^* is complementary to ω_e on ω. Set

$$\begin{aligned} A(\theta) &= A_e(\theta) + A_e^*(\theta), \\ H(\theta) &= H_e(\theta) + H_e^*(\theta), \end{aligned} \tag{4.2}$$

where $A_e(\theta)$ and $H_e(\theta)$ are integrals with the respective integrands of $A(\theta)$ and $H(\theta)$ and with ω_e as the domain of integration. The integrals $A_e^*(\theta)$ and $H_e^*(\theta)$ are similarly defined with ω_e^* as the domain of integration. The integral $A_e^*(\theta)$ is finite. Hence $A_e(\theta)$ is finite by virtue of our hypothesis that $A(\theta)$ is finite. We continue by showing that $H_e(\theta)$ is finite.

Recall that

$$\sum_i [g_i'(s)]^2 \equiv 1 \qquad (i = 1, \cdots, n).$$

Let η be an arbitrary small positive constant between 0 and 1 exclusive. If n numbers s_i all lie on a sufficiently short subinterval of the s-axis, then

$$\sum_i [g_i'(s_i)]^2 \geqq 1 - \eta > 0.$$

Now

$$A_e(\theta) = \frac{1}{16\pi} \int\int_{\omega_e} \sum_i \left| \frac{g_i'(s_i)[\theta(\alpha) - \theta(\beta)]}{\sin \dfrac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta \tag{4.3}$$

where s_i lies between $\theta(\alpha)$ and $\theta(\beta)$ and depends on α and β. If e is sufficiently small $|\theta(\alpha) - \theta(\beta)|$ will be arbitrarily small provided (α, β) is on ω_e. Hence for e sufficiently small

$$A_e(\theta) \geqq (1 - \eta) H_e(\theta). \tag{4.4}$$

But $A_e(\theta)$ is finite so that $H_e(\theta)$ will then be finite. We conclude that $H(\theta)$ is finite. The proof of the theorem is complete.

We note that $H_e(\theta) = 0$ for no admissible θ. We shall need the following lemma.

LEMMA 4.1. *The functions $\varphi(\alpha)$ are equicontinuous for φ on any compact subset A of Ω.* Cf. [5], p. 692.

Let $\varphi_1, \cdots \varphi_m$ be a finite set of points on A such that the e-neighborhoods of the respective points φ_k cover all of A. There exists a positive constant δ so small that the condition

$$|\alpha - \beta| < \delta \tag{4.5}$$

implies that

$$|\varphi_k(\alpha) - \varphi_k(\beta)| < e \qquad (k = 1, \cdots, m). \tag{4.6}$$

If however $\varphi(\alpha)$ is any point on A there exists a point $\varphi_j(\alpha)$ such that

$$|\varphi(\alpha) - \varphi_j(\alpha)| < e \tag{4.7}$$

for each value of α. From (4.6) and (4.7) we conclude that

$$|\varphi(\alpha) - \varphi(\beta)| < 3e$$

subject to (4.5). Hence the functions $\varphi(\alpha)$ of the set A are equicontinuous. With the aid of this lemma we can prove the following theorem.

THEOREM 4.2. *Corresponding to any subset B of Ω for which $A(\varphi)$ is bounded, the ratio of $A_e(\varphi)$ to $H_e(\varphi)$ tends uniformly to 1 as e tends to 0. Moreover when $H(\varphi)$ is bounded $A(\varphi)$ is bounded and conversely.*

By virtue of the condition of bounded compactness, B is a subset of a compact set A of Ω. According to the preceding lemma the functions $\varphi(\alpha)$ on B are equicontinuous. If then the function $\theta(\alpha)$ in (4.4) ranges over the set B, relation (4.4) holds, with η independent of the choice of θ on B and tending to 0 with e. We also note that the relation

$$A_e(\varphi) \leqq H_e(\varphi) \tag{4.8}$$

holds as a consequence of (4.0). The first statement in the theorem follows from (4.4) and (4.8). The second statement in the theorem follows similarly from (4.4), Lemma 4.1, and (4.1).

It is of interest to note that $H(\theta)$ may be infinite. In fact $H(\theta)$ is infinite if

$$\theta(\alpha) = (4\pi^2\alpha)^{\frac{1}{2}} \qquad 0 \leqq \alpha \leqq 2\pi.$$

We note that

$$\theta(0) = 0, \qquad \theta(2\pi) = 2\pi.$$

The definition of $\theta(\alpha)$ can be extended by requiring that

$$\theta(\alpha + 2\pi) = \theta(\alpha) + 2\pi.$$

In general $\theta(\alpha)$ will not satisfy the three point condition, but a suitable conformal transformation of the disc $r \leqq 1$ will lead to such a function, and for this transformed function ψ, $H(\psi)$ will again be infinite. Before coming to the condition of regularity at infinity we shall introduce a basic deformation.

The deformation $\Delta(\psi)$. Corresponding to each point ψ of Ω we shall define a continuous deformation $\Delta(\psi)$ of Ω. In this deformation the point ψ will be fixed. The time t shall run from 0 to 1 inclusive and an arbitrary point φ on Ω shall be replaced at the time t by the point

$$\varphi^t(\alpha) = t\psi(\alpha) + (1 - t)\varphi(\alpha).$$

Observe that $\varphi^t(\alpha)$ satisfies the three point condition by virtue of the fact that φ and ψ satisfy this condition. Moreover $\varphi^0 = \varphi$, $\varphi^1 = \psi$, and φ^t has a distance from φ^τ given by

$$\varphi^t\varphi^\tau = |t - \tau|\,\varphi\psi. \tag{4.9}$$

It is also clear that ψ is fixed under this deformation. We must show that $\varphi^t(\alpha)$ is a point of Ω.

To that end we set

$$h(t) = H(\varphi^t) = \frac{1}{16\pi}\int\int_\omega \left|\frac{\varphi^t(\alpha) - \varphi^t(\beta)}{\sin\dfrac{(\alpha - \beta)}{2}}\right|^2 d\alpha\, d\beta, \tag{4.10}$$

and observe that

$$\begin{aligned} h(t) &= (1 - t)^2H(\varphi) + t^2H(\psi) \\ &\quad + \frac{2t(1 - t)}{16\pi}\int\int_\omega \frac{[\psi(\alpha) - \psi(\beta)][\varphi(\alpha) - \varphi(\beta)]}{\sin^2\dfrac{(\alpha - \beta)}{2}}\, d\alpha\, d\beta. \end{aligned} \tag{4.11}$$

Upon applying Schwarz's inequality to the middle term we find that

$$\begin{aligned} h(t) &\leqq t^2H(\psi) + 2t(1 - t)H^{\frac{1}{2}}(\varphi)H^{\frac{1}{2}}(\psi) + (1 - t)^2H(\varphi) \\ &= [tH^{\frac{1}{2}}(\psi) + (1 - t)H^{\frac{1}{2}}(\varphi)]^2. \end{aligned} \tag{4.12}$$

Hence $h(t)$ is finite and $\varphi^t(\alpha)$ is a point of Ω. We infer that

$$h(t) \leqq \text{max. } [H(\psi), H(\varphi)].$$

Similarly

$$H_e(\varphi^t) \leqq [tH_e^{\frac{1}{2}}(\psi) + (1 - t)H_e^{\frac{1}{2}}(\varphi)]^2.$$

We thus have the following lemma.

LEMMA 4.2. *Under the deformation* $\Delta(\psi)$ *the image* φ^t *of an arbitrary point* φ *of* Ω *varies so that at each moment* t

$$H_e(\varphi^t) \leqq \text{max. } [H_e(\varphi), H_e(\psi)] \qquad (0 \leqq t \leqq 1).$$

Under $\Delta(\psi)$ the space Ω is continuously deformed on itself into the point ψ. Hence the connectivities of Ω are all 0 except that $R_0 = 1$. Under $\Delta(\psi)$ a subset of Ω on which $H(\varphi)$ is bounded by a constant c, is deformed over a subset of points of Ω on which $H(\varphi^t)$ is at most the maximum of c and $H(\psi)$. An arbitrary subset of Ω is deformed into a subset of Ω on which $A(\varphi)$ is bounded; in particular into ψ. We summarize in the following theorem.

THEOREM 4.3. *The space Ω is continuously deformable on itself into a point, and with the function $A(\varphi)$ satisfies the condition of regularity at infinity.*

5. The upper-reducibility of $A(\varphi)$

We shall show that $A(\varphi)$ is weakly upper-reducible at an arbitrary point ψ of Ω. Cf. §2.

To that end let $a = A(\psi)$, and let c be an arbitrary constant such that $c > a$. Let N_c be a neighborhood of ψ on Ω relative to the subset below $c+$. We shall show that the deformation $\Delta(\psi)$ of §4 taken for

$$0 \leqq t \leqq \tfrac{1}{2}, \tag{5.1}$$

and applied to a sufficiently small neighborhood N_c of ψ, satisfies the conditions (i) and (ii) of the definition of weak upper-reducibility. As in (4.2) we set

$$\begin{aligned} A(\varphi) &= A_e(\varphi) + A_e^*(\varphi) \\ H(\varphi) &= H_e(\varphi) + H_e^*(\varphi). \end{aligned} \tag{5.2}$$

We shall investigate the behavior of $A(\varphi)$ under the deformation $\Delta(\psi)$. The behavior of $H(\varphi)$ under $\Delta(\psi)$ is relatively easy to determine. We shall give a series of lemmas describing the behavior of $H(\varphi)$ under $\Delta(\psi)$ and relating the behavior of $A(\varphi)$ to that of $H(\varphi)$.

Under the deformation $\Delta(\psi)$ the point φ of Ω is replaced at the time t by the point φ^t. We set

$$\begin{aligned} A(\varphi^t) &= a(t), \quad A_e(\varphi^t) = a(t, e), \quad A_e^*(\varphi^t) = a^*(t, e), \\ H(\varphi^t) &= h(t), \quad H_e(\varphi^t) = h(t, e), \quad H_e^*(\varphi^t) = h^*(t, e), \end{aligned} \tag{5.3}$$

so that

$$\begin{aligned} a(t) &= a(t, e) + a^*(t, e), \\ h(t) &= h(t, e) + h^*(t, e). \end{aligned} \tag{5.4}$$

Upon differentiating (4.11) with respect to t and making use of the Schwarz inequality, we find that when $0 \leqq t \leqq \frac{1}{2}$

$$\begin{aligned} h'(t) &\leqq 2tH(\psi) + 2(1 - 2t)H^{\frac{1}{2}}(\varphi)H^{\frac{1}{2}}(\psi) + 2(t - 1)H(\varphi) \\ &= 2(1 - t)[H^{\frac{1}{2}}(\varphi) + \frac{t}{1 - t} H^{\frac{1}{2}}(\psi)][H^{\frac{1}{2}}(\psi) - H^{\frac{1}{2}}(\varphi)]. \end{aligned} \tag{5.5}$$

Similar relations hold when $H(\psi)$, $H(\varphi)$, and $h(t)$ are replaced by

$$H_e(\psi), \quad H_e(\varphi), \quad h(t, e)$$

respectively. We thus have the following lemma.

LEMMA 5.1. *Under the deformation* $\Delta(\psi)$

$$h_i(t, e) \leqq H_e^{\frac{1}{2}}(\varphi)[H_e^{\frac{1}{2}}(\psi) - H_e^{\frac{1}{2}}(\varphi)] \qquad (0 \leqq t \leqq \tfrac{1}{2})$$

whenever $H_e(\varphi) > H_e(\psi)$.

We continue with the following lemma.

LEMMA 5.2. *For a fixed* e, *and for* $0 < t < 1$ *the functions*

$$a_i^*(t, e), \qquad a^*(t, e) - a^*(1, e)$$

tend to 0 *uniformly with respect to* t, *as* φ *tends to* ψ.

For brevity we set

$$u = \varphi^t(\alpha), \qquad v = \varphi^t(\beta) \tag{5.6}$$

and recall that for $\sigma = \omega_e^*$

$$a^*(t, e) = \frac{1}{16\pi} \int\int_\sigma \sum_i \left| \frac{g_i(u) - g_i(v)}{\sin \dfrac{(\alpha - \beta)}{2}} \right|^2 d\alpha \, d\beta .$$

Upon differentiating under the integral sign with respect to t, we find that

$$a_i^*(t, e) = \frac{1}{8\pi} \int\int_\sigma \sum_i \left[\frac{g_i(u) - g_i(v)}{\sin^2 \dfrac{(\alpha - \beta)}{2}} \right] [u_t g_i'(u) - v_t g_i'(v)] \, d\alpha \, d\beta , \tag{5.7}$$

where $\sigma = \omega_e^*$ and

$$u_t = \psi(\alpha) - \varphi(\alpha), \qquad v_t = \psi(\beta) - \varphi(\beta). \tag{5.8}$$

The integral (5.7) is proper, and by virtue of (5.8) tends to 0 uniformly with respect to t as φ tends to ψ. Here e and ψ are fixed. Upon integrating $a_i^*(t, e)$ with respect to t between 1 and t, it follows that

$$a^*(t, e) - a^*(1, e)$$

tends to 0 uniformly with respect to t as φ tends to ψ. The proof of the lemma is complete.

In analyzing the behavior of $A(\varphi)$ under $\Delta(\psi)$ we shall need a formula for $a_i(t, e)$. To that end we shall make use of functions $a_i(u, v)$ which will be defined in (5.10).

The functions $a_i(u, v)$. We set

$$g_i(u) - g_i(v) = a_i(u, v)(u - v) \qquad (i = 1, \cdots, n), \tag{5.9}$$

where

$$a_i(u, v) = \int_0^1 g_i'[u + \tau(v - u)]d\tau. \tag{5.10}$$

In particular

$$a_i(u, u) \equiv g_i'(u)$$

so that

$$\sum_i a_i^2(u, u) \equiv 1. \tag{5.11}$$

It follows from (5.9) that $a_i(u, v)$ is of class C' when $u \neq v$. Recall that the functions $g_i'(s)$ possess a Lipschitz constant K. Upon using (5.10) we find that

$$\begin{aligned} |\, a_i(u + \Delta u, v + \Delta v) - a_i(u, v)\,| &\leqq K\,|\,\Delta u\,| \int_0^1 (1 - \tau)\, d\tau \\ &\quad + K\,|\,\Delta v\,| \int_0^1 d\tau \leqq K\{\,|\,\Delta u\,| + |\,\Delta v\,|\}. \end{aligned} \tag{5.12}$$

Thus K is a Lipschitz constant for each of the functions $a_i(u, v)$.

We shall need the t-derivative $(0 < t < 1)$ of the function obtained from $a_i(u, v)$ upon making the substitution (5.6). Recall the relations (5.8). Suppose first that

$$\varphi^t(\alpha) \neq \varphi^t(\beta) \tag{5.13}$$

for a given triple (t, α, β). Then $u \neq v$. Making use of the fact that $a_i(u, v)$ is of class C' in u and v when $u \neq v$, we see that the t-derivative of $a_i(u, v)$ exists and satisfies the condition

$$\left|\frac{d}{dt}[a_i(u, v)]\right| \leqq K\,|\,\varphi(\alpha) - \varphi(\beta)\,| + K\,|\,\psi(\alpha) - \psi(\beta)\,| \quad (0 < t < 1). \tag{5.14}$$

If on the other hand

$$\varphi^t(\alpha) = \varphi^t(\beta) \tag{5.15}$$

for a given triple (t, α, β), then

$$\varphi(\alpha) - \varphi(\beta) = \psi(\alpha) - \psi(\beta) = 0 \tag{5.16}$$

as one readily sees. Hence $u_t = v_t = 0$ for the set (t, α, β). Making use of (5.12) we see that the t-derivative of $a_i(u, v)$ again exists, in fact is 0, and in particular satisfies (5.14).

Moreover the t-derivative, $0 < t < 1$, of $a_i(u, v)$ is continuous in (t, α, β), subject to (5.6). This is at once clear in the case (5.13), since $a_i(u, v)$ is of class C' in u and v when $u \neq v$. But when (5.15) holds, (5.16) also holds and the t-derivative of $a_i(u, v)$ is continuous by virtue of (5.14).

A representation of $a_i(t, e)$. Let η be an arbitrarily small positive constant

between 0 and e, and let ω_e^η be the subset of ω_e on which

$$\eta \leqq |\alpha - \beta|.$$

For any subset σ of ω set

$$(5.17)' \qquad I(\sigma) = \frac{1}{16\pi} \int\int_\sigma \sum_i a_i^2(u, v) \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta,$$

$$(5.17)'' \qquad I_1(\sigma) = \frac{1}{8\pi} \int\int_\sigma \sum_i a_i(u, v) \frac{d}{dt} a_i(u, v) \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta,$$

$$(5.17)''' \qquad I_2(\sigma) = \frac{1}{16\pi} \int\int_\sigma \sum_i a_i^2(u, v) \frac{d}{dt} \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta,$$

with u and v given by (5.6). We note that

$$I(\omega_e) = a(t, e).$$

We shall obtain $a_i(t, e)$ by differentiating under the integral sign in $I(\omega_e)$. To justify this we first replace ω_e by ω_e^η, and note that for the resulting integral

$$\frac{d}{dt} I(\omega_e^\eta) = I_1(\omega_e^\eta) + I_2(\omega_e^\eta).$$

We shall show that the integrals

$$I_1(\omega_e^\eta), \qquad I_2(\omega_e^\eta) \qquad (0 < t < 1)$$

converge uniformly with respect to t to the integrals $I_1(\omega_e)$ and $I_2(\omega_e)$ as η tends to 0.

This is clear for I_1 since $d[a_i(u, v)]/dt$ is bounded in t, α, β, and the integral

$$\int\int_\omega \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta \qquad (0 < t < 1)$$

is bounded. It is also clear for I_2 upon observing that

$$\frac{d}{dt} \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 = \left[\frac{1}{\sin^2 \frac{(\alpha - \beta)}{2}} \right]$$

$$\cdot \{2t[\psi(\alpha) - \psi(\beta)]^2 + 2(1 - 2t)[\psi(\alpha) - \psi(\beta)][\varphi(\alpha) - \varphi(\beta)]$$

$$+ 2(t - 1)[\varphi(\alpha) - \varphi(\beta)]^2\}$$

so that the integral

$$(5.18) \qquad \int\int_\omega \left| \frac{d}{dt} \left| \frac{u - v}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 \right| d\alpha\, d\beta \qquad (0 < t < 1)$$

is bounded. We can accordingly differentiate $I(\omega_e)$ with respect to t under the integral sign, and so obtain the desired formula

$$a_t(t, e) = I_1(\omega_e) + I_2(\omega_e) \qquad (0 < t < 1). \tag{5.19}$$

We come to a basic lemma.

LEMMA 5.3. *For a fixed e and bounded $A(\varphi)$ the difference*

$$a_t(t, e) - h_t(t, e) \qquad (0 < t < 1)$$

tends to 0 uniformly with respect to t as φ tends to ψ.

Subject to (5.6)

$$h_t(t, e) = \frac{1}{16\pi} \int\int_{\omega_e} \frac{d}{dt} \left| \frac{u - v}{\sin \dfrac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta.$$

It follows from (5.19) upon setting

$$w = \sum_i a_i^2(u, v),$$

that, for $0 < t < 1$,

$$a_t(t, e) - h_t(t, e) = I_1(\omega_e) + \frac{1}{16\pi} \int\int_{\omega_e} (w - 1) \frac{d}{dt} \left| \frac{u - v}{\sin \dfrac{(\alpha - \beta)}{2}} \right|^2 d\alpha\, d\beta. \tag{5.20}$$

The integral $I_1(\omega_e)$ tends to 0 uniformly with respect to t as φ tends to ψ. For $d[a_i(u, v)]/dt$ tends to 0 uniformly with respect to t as φ tends to ψ in accordance with (5.14). The second integral in (5.20) tends to 0 uniformly with respect to t as φ tends to ψ since $w - 1$ tends to 0 as φ tends to ψ, and since the integral (5.18) is bounded when $A(\varphi)$ is bounded. The lemma follows from (5.20).

Recall that

$$a_t(t) = a_t(t, e) + a_t^*(t, e) \qquad (0 < t < 1)$$

and that $a_t^*(t, e)$ tends to 0 with $\psi\varphi$ in accordance with Lemma 5.2. But

$$a_t(t) - h_t(t, e) = [a_t(t, e) - h_t(t, e)] + a_t^*(t, e).$$

The preceding lemma accordingly leads to the following.

LEMMA 5.4. *For a fixed e and bounded $A(\varphi)$ the difference*

$$a_t(t) - h_t(t, e) \qquad (0 < t < 1) \tag{5.21}$$

together with the integral

$$[a(t) - a(0)] - [h(t, e) - h(0, e)] \tag{5.22}$$

tends to 0 uniformly with respect to t as φ tends to ψ.

The principal theorem of this section is the following.

THEOREM 5.1. *The function $A(\varphi)$ is weakly upper-reducible at each point ψ of the space Ω.*

As previously noted we set $A(\psi) = a$ and let c be an arbitrary constant such that $c > a$. Let N_c be a neighborhood of ψ relative to $A(\varphi) \leqq c$. We shall show that the neighborhood N_c, suitably restricted, together with the deformation $\Delta(\psi)$ with $0 \leqq t \leqq \frac{1}{2}$, will satisfy the conditions (i) and (ii) of the definition of weak upper-reducibility. Cf. §2.

Let r be a constant between 0 and 1 exclusive, such that

$$r^2 c > a. \tag{5.23}$$

We consider only those points φ for which $A(\varphi) \leqq c$. For such points φ there exists a positive constant e, in accordance with Theorem 4.2, such that

$$A_e(\varphi) > rH_e(\varphi). \tag{5.24}$$

We hold this value of e fast. Note that rc lies between c and a. Let N' be the subset of N_c on which

$$A(\varphi) \geqq rc \tag{5.25}$$

and N'' be the complement of N' on N_c so that

$$N_c = N' + N''.$$

We shall impose three conditions on N_c.

First recall that

$$A(\varphi) = A_e(\varphi) + A_e^*(\varphi). \tag{5.26}$$

Set

$$X = H_e(\varphi) - H_e(\psi).$$

By virtue of our choice of e

$$X > A_e(\varphi) - r^{-1}A_e(\psi).$$

Making use of (5.26) we find that

$$X > A(\varphi) - r^{-1}A(\psi) - A_e^*(\varphi) + r^{-1}A_e^*(\psi).$$

Upon replacing the last term by the smaller term $A_e^*(\psi)$ and using (5.25) we see that on N'

$$X > \frac{r^2 c - a}{r} - [A_e^*(\varphi) - A_e^*(\psi)]. \tag{5.27}$$

The first condition on N_c is the following.

We suppose N_c so small that on N', X is positive and bounded from 0.

This is possible by virtue of (5.27). For $r^2 c - a > 0$ in accordance with (5.23), and $A_e^*(\varphi)$ tends to $A_e^*(\psi)$ as φ tends to ψ.

We turn to conditions (i) and (ii) of the definition of weak upper-reducibility.

Condition (i). Suppose that

$$\varphi \subset N' \qquad 0 < t \leqq \tfrac{1}{2}. \tag{5.28}$$

Then the difference

$$X = H_e(\varphi) - H_e(\psi)$$

is positive and bounded from 0. It follows from Lemma 5.1 that $h_t(t, e)$ is negative and bounded from 0. By virtue of Lemma 5.4, N_c can be restricted further so that when (5.28) holds, $a_t(t)$ is likewise negative and bounded from 0. Under $\Delta(\psi)$, N' will then be deformed with a displacement function [cf. (4.9)] into a set below $c - \eta$, where η is a suitably chosen positive constant. But the constant η can be chosen arbitrarily small, in particular so small that

$$rc < c - \eta.$$

The subset of N_c not below $c - \eta$ lies on N', and will accordingly be deformed into a set below $c - \eta$ Condition (i) is accordingly satisfied.

Condition (ii). We shall show that condition (ii) is satisfied by $\Delta(\psi)$ provided N_c be restricted still further. We consider two cases

CASE A. $H_e(\varphi) > H_e(\psi)$. It follows from Lemma 4.2 that $H_e(\varphi)$ is not increased beyond its initial value. It follows from Lemma 5.4 that $A(\varphi)$ will be increased beyond its initial value at most by an amount which tends to 0 as φ tends to ψ. Recall that N'' lies on the subset

$$A(\varphi) < rc < c - \eta$$

of Ω. We can accordingly suppose N_c so small that N'' is deformed below $c - \eta$ throughout $\Delta(\psi)$, at least in Case A.

CASE B. $H_e(\varphi) \leqq H_e(\psi)$. We shall show that $a(t)$ remains less than rc and hence less than $c - \eta$, provided N_c is sufficiently small. Recall that

$$\begin{aligned} a(t) &= a(t, e) + a^*(t, e) \\ &\leqq h(t, e) + a^*(t, e). \end{aligned}$$

In this case

$$h(t, e) \leqq H_e(\psi)$$

by virtue of Lemma 4.2. Hence

$$\begin{aligned} a(t) &\leqq H_e(\psi) + a^*(t, e) \leqq r^{-1}A_e(\psi) + a^*(t, e) \\ &\leqq r^{-1}A_e(\psi) + A_e^*(\psi) + [a^*(t, e) - A_e^*(\psi)] \\ &\leqq r^{-1}[A_e(\psi) + A_e^*(\psi)] + [a^*(t, e) - a^*(1, e)] \\ &\leqq ar^{-1} + [a^*(t, e) - a^*(1, e)]. \end{aligned}$$

Hence

$$a(t) - cr \leqq \left[\frac{a - cr^2}{r}\right] + [a^*(t, e) - a^*(1, e)].$$

The first bracket is negative by virtue of the choice of r, while the second bracket will be less than half the first bracket in absolute value if N_c is sufficiently

small. Cf. Lemma 5.2. For such a choice of N_c, φ will be deformed under $\Delta(\psi)$ so as to remain below cr, at least in Case B.

If the neighborhood N_c is subjected to the above three conditions, condition (ii) will be satisfied by $\Delta(\psi)$ for $0 \leqq t \leqq \frac{1}{2}$. For points of Ω on N_c which are initially below $c - \eta$ are either on N' or on N''. If on N' they are deformed under $\Delta(\psi)$ so as to remain below $c - \eta$, and if on N'' they remain below $c - \eta$ in accordance with the results in Cases A and B. The condition (ii) is thus satisfied. The proof of the theorem is complete.

6. The homotopy theorem

We have shown that the space Ω and the function $A(\varphi)$ satisfy the three conditions of bounded compactness, regularity at infinity, and weak upper-reducibility. The general function theory of §2 then applies and implies existential relations between the homotopic critical points of $A(\varphi)$. It remains to show that a homotopic critical point φ defines a minimal surface bounded by g. This is the homotopy theorem.

The harmonic surface $x_i = y_i(u, v)$ of §2 will be said to be *minimal* if the parameters u, v are isothermal. In the notation of differential geometry the parameters u, v are isothermal if

$$E \equiv G, \qquad F \equiv 0 \qquad (r < 1).$$

A point φ of Ω which defines a minimal surface will be termed *differentially critical*, as will the p-curve

$$x_i = g_i[\varphi(\alpha)] \tag{6.1}$$

defined by φ. A point φ or p-curve which is not differentially critical will be termed *differentially ordinary*.

Let $\varphi(\alpha)$ be a transformation which satisfies the conditions imposed on points of Ω with the possible exception of the three point condition. Suppose however that the values

$$\varphi(\alpha_k) \qquad (k = 1, 2, 3)$$

are distinct. The transformation φ as well as the p-curve defined by φ will then be said to satisfy the *weak three point condition*. The space of such transformations φ will be denoted by Ω' and the space of p-curves defined by points φ of Ω' will be denoted by Z. Distances in the space Z are to be defined as in §2. Two curves of Z are to be regarded as identical if and only if their distance is null. By definition a curve

$$x_i = g_i[\psi(\alpha)] = p_i(\alpha)$$

of Z is "defined" by the point $\psi(\alpha)$ of Ω'. If this curve is defined by a second point $\theta(\alpha)$ of Ω' one readily sees that

$$\theta(\alpha) - \psi(\alpha) \equiv 0 \pmod{2\pi}. \tag{6.2}$$

Two points $\varphi(\alpha)$ and $\psi(\alpha)$ of Ω' will be said to be in the same *conformal class* if there exists a conformal transformation T of the disc $r \leqq 1$ into itself such that the point on the circle $r = 1$ with angular coördinate $\varphi(\alpha)$ is carried into the point on $r = 1$ with angular coördinate $\psi(\alpha)$ for each α.

Given a curve $[p]$ of Z, a point $\varphi = f(p)$ of Ω will be defined as follows. Let $\psi(\alpha)$ be a point of Ω' defining $[p]$. We take $f[p]$ as that point $\varphi(\alpha)$ of Ω which is in the conformal class of $\psi(\alpha)$. If $\theta(\alpha)$ also defines $[p]$, it is clear from (6.2) that $\varphi(\alpha)$ is in the conformal class of $\theta(\alpha)$. Thus $f[p]$ is uniquely determined by $[p]$. It is also seen that $f[p]$ varies continuously on Ω with $[p]$ on Z. We have in $f[p]$ a continuous mapping of Z into Ω.

The Dirichlet function $D(p)$ is defined for $[p]$ on Z. Relative to $D(p)$ and Z the term homotopically ordinary is well defined. We shall show that when the harmonic surface defined by $[q]$ is not minimal, $[q]$ is homotopically ordinary. We shall obtain the corresponding result for $A(\varphi)$ with the aid of the following lemma.

LEMMA 6.1. *If a point ψ of Ω defines a point $[q]$ of Z which is homotopically ordinary relative to $D(p)$, then ψ is homotopically ordinary relative to $A(\varphi)$.*

If $[q]$ is homotopically ordinary relative to $D(p)$ there exists a D-deformation Λ on Z of a neighborhood of $[q]$ relative to the subset $D(p) \leqq D(q)$, such that $[q]$ is actually displaced. We can transform the deformation Λ of Z into a deformation Λ^* of Ω, mapping each point $[p]$ of Z onto the point $f[p]$ of Ω. Since

$$D(p) = A\{f[p]\},$$

it appears at once that Λ^* is an A-deformation of a neighborhood of ψ relative to the subset $A \leqq A(\psi)$, and that ψ is actually displaced. The proof of the lemma is complete.

A deformation Λ on Z. Let $x_i = p_i(\alpha)$ be a curve of Z. We shall consider a one-parameter family of transformations of the parameter α into a parameter μ of the form[1]

$$\mu = \alpha + e\lambda(\alpha), \tag{6.3}$$

where $\lambda(\alpha)$ is a real function of class C''' with a period 2π. The number e is constant in any one transformation, and is restricted to a range

$$|e| \leqq e_0 \qquad (e_0 > 0) \tag{6.4}$$

so small that

$$1 + e\lambda'(\alpha) \neq 0. \tag{6.5}$$

For such values of e the transformation (6.3) has a single-valued inverse

$$\alpha = m(\mu, e) \tag{6.6}$$

[1] A variation of the form (6.3) in which λ is a sine or a cosine is used by Douglas in [1] p. 289. The reader will note the similarity of our first transformations to those of Douglas. However we are led to properties of uniformity unnecessary for Douglas' purposes.

with

$$\alpha \equiv m(\alpha, 0).$$

We introduce the p-curve

$$x_i = p_i[m(\mu, e)] \tag{6.7}$$

with parameter μ. For $e = 0$ this p-curve satisfies the weak three point condition. If e_0 is sufficiently small the weak three point condition will still be satisfied for $|e| \leqq e_0$. We suppose e_0 so chosen. The constant e_0 thus depends upon $\lambda(\alpha)$ and upon $[p]$. The curve (6.7) represents a point on Z provided the Dirichlet integral $d(p, e)$ defined by the curve (6.7) is finite. We shall show that $d(p, e)$ is finite.

In representing the integral $d(p, e)$ we shall be led to investigate the function

$$w(\alpha, \beta, e) = \sin\left[\frac{\alpha - \beta + e\{\lambda(\alpha) - \lambda(\beta)\}}{2}\right].$$

This function appears in the denominator of the integrand of $d(p, e)$. We note that

$$\alpha - \beta + e[\lambda(\alpha) - \lambda(\beta)] = (\alpha - \beta)[1 + eg(\alpha, \beta)], \tag{6.8}$$

where $g(\alpha, \beta)$ is of class C''. The function $w(\alpha, \beta, e)$ accordingly has the form

$$\sin\left[\frac{(\alpha - \beta) + e\{\lambda(\alpha) - \lambda(\beta)\}}{2}\right] = \sin\frac{(\alpha - \beta)}{2}[1 + eh(\alpha, \beta, e)]. \tag{6.9}$$

The function $h(\alpha, \beta, e)$ is to be regarded as defined by (6.9), except of course when $e = 0$ and when $\alpha - \beta \equiv 0 \bmod 2\pi$. For these exceptional values we see that $h(\alpha, \beta, e)$ can be so defined as to be continuous.

We shall show that

$$1 + eh(\alpha, \beta, e) \neq 0 \qquad |e| \leqq e_0. \tag{6.10}$$

To that end we observe that $w(\alpha, \beta, e) = 0$ when and only when

$$(\alpha - \beta) + e[\lambda(\alpha) - \lambda(\beta)] \equiv 0 \pmod{2\pi}. \tag{6.11}$$

When (6.11) holds for a pair (α, β) there will exist a pair (α_1, β_1) such that

$$0 \leqq \alpha_1 < 2\pi, \qquad 0 \leqq \beta_1 < 2\pi,$$

$$\alpha \equiv \alpha_1, \qquad \beta \equiv \beta_1 \pmod{2\pi}$$

and such that

$$\alpha_1 + e\lambda(\alpha_1) = \beta_1 + e\lambda(\beta_1).$$

Since the transformation (6.3) is one-to-one we conclude that $\alpha_1 = \beta_1$. Hence (6.11) holds only when $\alpha - \beta \equiv 0 \bmod 2\pi$. Condition (6.10) now follows from (6.9).

We observe that $h(\alpha, \beta, e)$ has the period 2π in α and β, and with the aid of

(6.8) and (6.9) we see that $h(\alpha, \beta, e)$ is of class C''. Upon differentiating the respective members of (6.9) with respect to β and setting $\beta = \alpha$ we find that

$$\lambda'(\alpha) = h(\alpha, \alpha, e). \tag{6.12}$$

Upon differentiating (6.9) twice with respect to β and setting $\beta = \alpha$ we find that

$$\lambda''(\alpha) = 2h_\beta(\alpha, \alpha, e). \tag{6.13}$$

We return to the integral

$$d(p, e) = \frac{1}{16\pi} \int \int_\omega \sum_i \left| \frac{p_i[m(\mu, e)] - p_i[m(\nu, e)]}{\sin \frac{(\mu - \nu)}{2}} \right|^2 d\mu \, d\nu. \tag{6.14}$$

Setting

$$\mu = \alpha + e\lambda(\alpha), \qquad \nu = \beta + e\lambda(\beta)$$

and taking α and β as the variables of integration in (6.14), this integral takes the form

$$d(p, e) = \frac{1}{16\pi} \int \int_\omega \sum\nolimits_i \left| \frac{p_i(\alpha) - p_i(\beta)}{w(\alpha, \beta, e)} \right|^2 [1 + e\lambda'(\alpha)][1 + e\lambda'(\beta)] \, d\alpha \, d\beta. \tag{6.15}$$

From the representation (6.9) of $w(\alpha, \beta, e)$, from (6.10), and the finiteness of $D(p)$, we infer that $d(p, e)$ is finite, in fact bounded for $| e | \leqq e_0$.

Making use of the binomial theorem we see that

$$\frac{1}{[1 + eh(\alpha, \beta, e)]^2} = 1 - 2eh(\alpha, \beta, 0) + e^2 x(\alpha, \beta, e),$$

where $x(\alpha, \beta, e)$ is of class C'. We can thus write $d(p, e)$ in the final form

$$d(p, e) = D(p) + eS(p) + e^2 R(p, e) \qquad | e | \leqq e_0, \tag{6.16}$$

where

$$S(p) = \frac{1}{16\pi} \int \int_\omega \sum_i \left| \frac{p_i(\alpha) - p_i(\beta)}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 [\lambda'(\alpha) + \lambda'(\beta) - 2h(\alpha, \beta, 0)] \, d\alpha \, d\beta, \tag{6.17}$$

$$R(p, e) = \frac{1}{16\pi} \int \int_\omega \sum_i \left| \frac{p_i(\alpha) - p_i(\beta)}{\sin \frac{(\alpha - \beta)}{2}} \right|^2 z(\alpha, \beta, e) \, d\alpha \, d\beta, \tag{6.18}$$

with $z(\alpha, \beta, e)$ of class C'.

Observe that $S(p)$ is independent of e. It follows from (6.12) and (6.13) that

$$\lambda'(\alpha) + \lambda'(\beta) - 2h(\alpha, \beta, 0) = (\alpha - \beta)^2 k(\alpha, \beta), \tag{6.19}$$

where $k(\alpha, \beta)$ is continuous in α and β. From (6.19) we infer that the singularity in the integrand of $S(p)$ when $\alpha = \beta$ is removable. We conclude that $S(p)$ varies continuously with $[p]$.

For $|e| \leqq e_0$ the functions $z(\alpha, \beta, e)$ and $z_e(\alpha, \beta, e)$ admit a finite upper bound, $\kappa > 0$, independent of $[p]$. From (6.18) we see that

$$|R(p, e)| \leqq \kappa D(p), \qquad |R_e(p, e)| \leqq \kappa D(p). \tag{6.20}$$

We summarize in the following lemma.

LEMMA 6.2. *Let $[p(\alpha)]$ be an admissible representation of g for which $D(p)$ is finite, and let $\lambda(\alpha)$ be a function of α of class C''' with period 2π. If e_0 is a sufficiently small positive constant and $|e| \leqq e_0$, the transformation $\mu = \alpha + e\lambda(\alpha)$ has a single-valued inverse $\alpha = m(\mu, e)$ and the Dirichlet integral $d(p, e)$ corresponding to the boundary curve*

$$x_i = p_i[m(\mu, e)] \tag{6.21}$$

has the form

$$d(p, e) = D(p) + eS(p) + e^2R(p, e) \qquad |e| \leqq e_0, \tag{6.22}$$

where the variation $S(p)$ depends continuously on $[p]$, and the functions $R(p, e)$ and $R_e(p, e)$ are bounded as shown in (6.20).

We can now establish the homotopy theorem in the space Z.

THEOREM 6.1. *A curve $[q]$ of Z which is differentially ordinary relative to $D(p)$ is homotopically ordinary relative to $D(p)$.*

We are assuming that the harmonic surface defined by $[q]$ is not isothermally represented by its parameters (u, v). According to Douglas, cf. [1], pp. 297–298, there then exists an analytic function $\lambda(\alpha)$ with period 2π in α such that the Dirichlet integral $d(q, e)$ of Lemma 6.2 has a derivative $d_e(q, 0)$ which is negative. We subject the parameter α of the p-curves neighboring $[q]$ to the transformation

$$\mu = \alpha' + e\lambda(\alpha) \qquad |e| \leqq e_0, \tag{6.23}$$

holding e fast. As previously, we suppose e_0 so small that $1 + e\lambda' \neq 0$. If N is a sufficiently small neighborhood of $[q]$ on Z and e_0 is sufficiently small, the curves which result upon applying the transformation (6.23) to the parameter α of p-curves of N, will satisfy the weak three point condition. We suppose N and e_0 so chosen. If $d(p, e)$ is the Dirichlet integral for the curve thereby replacing $[p]$, (6.22) now holds, as well as the relations (6.20).

Set $D(q) = c$. We have the relations

$$S(p) = d_e(p, 0) \qquad S(q) = d_e(q, 0) < 0.$$

For points $[p]$ below c the functions $R(p, e)$ and $R_e(p, e)$ are bounded by κc by virtue of (6.20). It follows from (6.22) that for $[p]$ on a sufficiently small neighborhood N_0 of $[q]$ relative to $D(p) \leqq c$, and for $|e|$ at most a sufficiently small positive constant $e_1 < e_0$, the derivative $d_e(p, e)$ is negative and bounded from 0.

The deformation $\Lambda(\lambda)$. Corresponding to the above function $\lambda(\alpha)$ let $\Lambda(\lambda)$ be a deformation of curves $[p]$ of Z in which each curve $[p]$ is replaced at the time e by the curve

$$x_i = p_i[m(\mu, e)] \qquad (0 \leqq e \leqq e_1).$$

This deformation is clearly a D-deformation of the neighborhood N_0, restricted as above. Under $\Lambda(\lambda)$, $[q]$ is displaced, since $d(q, e)$ is a decreasing function of e. Hence q is homotopically ordinary and the proof of the theorem is complete.

Lemma 6.1 together with the preceding theorem give us the fundamental theorem of this section.

THEOREM 6.2. *A homotopic critical point of $A(\varphi)$ is differentially critical and so defines a minimal surface bounded by g.*

We add the following fact of interest. If a point $\varphi(\alpha)$ on Ω defines a minimal surface, the transformation $\theta = \varphi(\alpha)$ is *strictly increasing*. This has been established by Douglas, Radó and others. A particularly simple proof is given by Radó [5], p. 75.

7. The removal of finiteness conditions

We return to the general function-theory of §2. We shall see that the condition of Theorem 2.1 of §2 that the homotopic critical points of F on M be finite in number implies the condition of Theorem 2.1 that the type numbers of the critical points be finite, at least for the case of the functional $A(\varphi)$ on Ω. That this is not true of F in general is shown by the following example.

EXAMPLE. The space M shall be a subset of the x, y, z space with $z \geqq 0$. It shall consist of the set of all points on circles with diameters coinciding with the segment $0 \leqq z \leqq 1$ of the z-axis and with tangent directions at the origin which form a perfect nowhere dense set. The function $F(p)$ shall equal the value of z at p. There are just two homotopic critical points, the origin and the point (0, 0, 1). The latter point has an infinite type number M_1. We also note that R_1 is infinite, in fact has the power of the continuum.

Peculiarities of this sort are excluded if in addition to conditions I, II, and III of §2 we add condition IV as follows.

IV. *Local F-connectedness of M.* Let p be a point of M at which $F(p) = c$. The set M will be said to be locally *F-connected of order* $n \geqq 0$ at p, if corresponding to each positive constant δ there exists a positive constant η, such that each singular n-sphere on the δ-neighborhood of p and below $c + \delta$ bounds an n-cell of norm at most η below $c + \eta$. See Morse [7], p. 25. We shall presently show that Ω is A-connected of all orders. We state the following theorem.

THEOREM 7.1. *If conditions* I *to* IV *are satisfied by F and M, and if the homotopic critical points of F are finite in number, then the connectivities R_k of M and the type numbers of the critical points of M are likewise finite.*

We shall prove the following theorem.

THEOREM 7.2. *The functional $A(\varphi)$ is locally A-connected of each order n at each point ψ of Ω.*

Set $A(\psi) = c$. We refer to the deformation $\Delta(\psi)$ of §4 recalling that $\Delta(\psi)$ is linear in form, holds ψ fast and contracts Ω to ψ. It is then clear that an n-sphere on a sufficiently small neighborhood of ψ will bound on a preassigned neighborhood of ψ. Cf. (4.9). That part of the condition of local A-connectedness which concerns values of $A(\varphi)$ may be referred to the following lemma.

LEMMA 7.1. *Corresponding to a positive constant η there exists a positive constant δ such that the δ-neighborhood of ψ below $c + \delta$ is deformed under $\Delta(\psi)$ so as to remain below $c + \eta$.*

Let δ_0 be a positive constant less than η and let r be a positive constant less than 1 such that

$$c + \delta_0 < r(c + \eta). \tag{7.1}$$

A first condition on δ is that $\delta < \delta_0$. Let N_δ be the δ-neighborhood of ψ below $c + \delta$. We shall make use of the notation of §4. In accordance with Theorem 4.2 of §4 there exists a positive constant e so small that for points φ on N_δ

$$A_e(\varphi) > rH_e(\varphi). \tag{7.2}$$

We shall hold this value of e fast. We distinguish two cases.

CASE A. $H_e(\varphi) > H_e(\psi)$. With reference to the way φ is deformed under $\Delta(\psi)$ we see that

$$\begin{aligned} a(t) &= a(t, e) + a^*(t, e) \\ &\leqq h(t, e) + a^*(t, e). \end{aligned} \tag{7.3}$$

Referring to Lemma 4.2 we see that in this case

$$\begin{aligned} a(t) &\leqq H_e(\varphi) + a^*(0, e) + [a^*(t, e) - a^*(0, e)] \\ &\leqq r^{-1}[A_e(\varphi) + A_e^*(\varphi)] + [a^*(t, e) - a^*(0, e)] \\ &\leqq r^{-1}(c + \delta) + [a^*(t, e) - a^*(0, e)]. \end{aligned} \tag{7.4}$$

The first term on the right is less than $c + \eta$, and it follows from Lemma 5.2 that if δ is sufficiently small and φ is on N_δ then $a(t) < c + \eta$.

CASE B. $H_e(\varphi) \leqq H_e(\psi)$. The proof in this case is similar, replacing φ by ψ in (7.4) and setting $t = 1$ where we previously set $t = 0$. Continuing as before we find that $a(t) < c + \eta$. The proof of the lemma is complete. The theorem follows from the lemma as previously indicated.

The condition that the critical points be finite in number can be replaced by the much less restrictive condition that the critical values be isolated. To proceed we shall need several definitions. For further details and proof see Morse [8].

By the complete critical set ω at the level c is meant the set of all homotopic critical points at which $F = c$. It follows from the definition of a homotopic critical point that ω is closed relative to the set $F = c$. By a *critical set* σ at the level c is meant any subset of ω which is closed in ω and at a positive distance from $\omega - \sigma$. A neighborhood of σ which is at a positive distance from $\omega - \sigma$ will be termed *separate*. By the k^{th} *cap type number* of σ is meant the dimension m_k of a maximal group of k-caps with cap-limit c on any separate neighborhood of σ. This type number is independent of the particular separate neighborhood of σ which is employed. By the k^{th} *cap type number* M_k of F, is meant the sum of

the k^{th} cap type numbers of an ensemble of disjoint critical sets which cover the set of homotopic critical points of F. The number M_k is independent of the particular way into which the critical points are divided into disjoint critical sets. The number M_k may be infinite and as such is regarded as a cardinal number.

The value of F at a homotopic critical point will be termed a *homotopic critical value*. We state the following theorem.

THEOREM 7.3. *Suppose that F and M satisfy conditions* I *to* IV *and that the homotopic critical values cluster at most at infinity. The k^{th} cap type number m_k of each critical set is finite. If M_k and R_k are finite, and if we set $e_k = M_k - R_k$, then*

$$\begin{aligned} e_0 &\geqq 0, \\ e_1 - e_0 &\geqq 0, \\ e_2 - e_1 + e_0 &\geqq 0, \\ &\cdots\cdots\cdots\cdots, \end{aligned}$$

the number of relations in general being infinite.

When applied to minimal surfaces all the hypotheses of the theorem are fulfilled with the possible exception of the hypotheses that the critical values are isolated and that M_k is finite. If the hypotheses of the theorem are not fulfilled there must be an infinite set of critical values, since the numbers m_k are finite. In the minimal surface theory there must then be an infinite set of minimal surfaces with distinct areas or else the conditions of the theorem are fulfilled.

We shall now state a theorem from which all trace of conditions on the finiteness of the number of critical values is removed.

THEOREM 7.4. *Suppose F and M satisfy conditions* I *to* IV. *Let z be a class of relative k-cycles u with a modulus $F \leqq a$. Suppose the cycles u are mutually homologous but non-bounding mod $F \leqq a$, on $F \leqq b$ for some value $b > a$. There is then a least number $c > a$ such that $F \leqq a$ contains a cycle of z, and at the level c there is at least one homotopic critical point q such that each critical set which contains q has a positive k^{th} cap type number.*

In this theorem the number a may be less than the absolute minimum of F, in which case u is an ordinary cycle non-bounding on $F \leqq b$. If in addition b is infinite, u is non-bounding on M. In particular u may be a 0-cycle. If $R_0 = 1$ the number c in the theorem is then the absolute minimum of F. In the case of the minimal surface problem this implies the existence of a minimal surface of absolute minimum type. This result is stated merely to give an insight into the way the theorem works.

A critical set of points σ will be termed *minimizing* if there exists a neighborhood N of σ such that whenever p is a point of N not on σ, $F(p)$ exceeds the value of F on σ. With this understood we have the following corollary of the theorem.

COROLLARY 7.1. *Suppose that F and M satisfy conditions* I *to* IV *and that $R_0 = 1$. Corresponding to any two disjoint minimizing critical sets there exists at*

least one homotopic critical point p such that each critical set which contains p has a positive first type number m_1.

Both the theorem and the corollary are false if Vietoris cycles are replaced by singular cycles. The theorem and corollary apply at once to the minimal surface problem inasmuch as conditions I to IV are fulfilled by $A(\varphi)$ and Ω, and $R_0 = 1$ on Ω. The application of the theorem is not limited by the existence of infinitely many critical sets or critical values. But in such a case the conclusion that there exists another critical set would add nothing were it not stated that the first type number is positive. By virtue of the homotopy theorem each homotopic critical point of Ω determines a minimal surface bounded by g.

8. An example

We shall give an example of a closed curve $\gamma(e)$ which satisfies the conditions imposed on the closed curve g of §2, and corresponding to which the Douglas functional $A(\varphi)$ on Ω possesses two disjoint minimizing critical sets. Corollary 7.1 of §7 then applies and we can infer the existence of at least one minimal surface, defined by a homotopic critical point φ on Ω, such that every critical set which contains φ has a positive first type number.

Before defining $\gamma(e)$ we shall make certain preliminary remarks of a general nature. Let

$$x_i = x_i(u, v), \qquad x_i = y_i(u, v) \qquad (u^2 + v^2 \leqq 1) \tag{8.1}$$
$$(i = 1, \cdots, n)$$

be two harmonic surfaces bounded by g. We shall say that these surfaces are *conformally equivalent* if there exists a conformal map (direct or inverse)

$$u' = h(u, v),$$
$$v' = k(u, v),$$

of the disc $r \leqq 1$ onto itself such that

$$y_i[h(u, v), k(u, v)] \equiv x_i(u, v) \qquad (i = 1, \cdots, n).$$

The harmonic surface $S : x_i = x_i(u, v)$ will be said to *determine* a point φ on Ω if S is conformally equivalent to a harmonic surface defined by φ in the sense of §2. We conclude that a necessary condition that two harmonic surfaces of the form (8.1) *determine the same point* φ on Ω is that they be conformally equivalent.

Let Π

$$y_i = f_i(x_1, \cdots, x_n) \qquad (i = 1, \cdots, n)$$

represent a one-to-one continuous transformation of the space (x) onto itself. The surface S will be said to have the *image*

$$x_i = f_i[x_1(u, v), \cdots, x_n(u, v)]$$

under Π and be said to be *conformally invariant* under Π, if conformally equivalent to its image under Π. In particular if the image of a minimal surface S under Π is an admissible minimal surface S', S and S' will determine the same homotopic critical point φ on Ω only if S and S' are conformally equivalent.

As previously let S be a harmonic surface $x_i = x_i(u, v)$ bounded by g with $u^2 + v^2 \leqq 1$. Suppose g is met by an $(n-1)$-plane π in a set of m-points, $m > 1$. In the representation of S as the continuous image of the disc $r \leqq 1$ suppose these m points are the images of the points

$$q_1, \cdots, q_m$$

on the circle $r = 1$ in the (u, v)-plane. We suppose that g is not tangent to π at its points of intersection with g. We are concerned with the intersection of S with π. To that end we term a curve λ on the disc $r \leqq 1$ *canonical* if it possesses the following properties.

(1). It is the homeomorph of a closed line segment.

(2). Its end points lie on the circle $r = 1$ and its other points on the open disc $r < 1$.

(3). If (u_0, v_0) is an inner point of λ the points on λ neighboring (u_0, v_0) are representable in the form

$$u = u(t), \qquad v = v(t),$$

where $u(t)$ and $v(t)$ are real functions, analytic in t at $t = 0$, and defined for all sufficiently small values of $|t|$, where moreover

$$u_0 = u(0), \qquad v_0 = v(0), \qquad u'^2(t) + v'^2(t) \neq 0.$$

Canonical curves are thus real, simple, regular, and analytic. It follows from the theory of harmonic functions,[2] that the intersection of S with π is of the following nature.

The image on the disc $r \leqq 1$ of the intersection of S with π is made up of a finite number of canonical arcs with the following properties. The end points of these arcs lie at the points q_i on the circle $r = 1$. There is an odd number of these arcs terminating at each point q_i. Each of the regions into which these arcs divide the unit disc contains an arc of the circle $r = 1$ on its boundary. Two canonical arcs intersect in at most one point.

We come to the definition[3] of our space curve $\gamma(e)$. We begin by defining a curve $H(e)$ in the (x, y)-plane. Let e be a positive constant or zero. The curve $H(e)$ shall consist of the following three arcs joined in the order written.

$$y = 1 + e - \cos x \qquad (0 \leqq x \leqq \pi) \tag{8.2}$$

---

[2] The properties of canonical arcs here enumerated may be derived by considerations similar to those employed by Radó in [3] pp. 34–35. Further references are given by Radó.

[3] Curves with forms similar to ours have been used for various purposes. See Douglas [2] p. 208. Courant has demonstrated soap films bounded by curves of this shape. We are not aware of any previous proof of Theorem 8.1.

$$x = \pi + (2 + e)\cos\theta, \qquad y = (2 + e)\sin\theta \qquad \left(\frac{\pi}{2} > \theta > -\frac{\pi}{2}\right) \tag{8.3}$$

$$y = -(1 + e - \cos x) \qquad (0 \leqq x \leqq \pi). \tag{8.4}$$

The curve $H(e)$ has the appearance of a horseshoe magnet. The end points of $H(e)$ are the points $(0, e)$ and $(0, -e)$ on the y-axis. We term the points $(0, e)$ and $(0, -e)$ the *vertices* of $H(e)$ and the positive x-axis the *axis* of $H(e)$. The segment $x = 0$, $-e \leqq y \leqq e$, will be called the *base* of $H(e)$, but is not included in $H(e)$. We shall make use of congruent copies of $H(e)$, carrying over the above terms to these copies.

Let $\alpha(e)$ denote the area bounded by $H(e)$ and its base. Let $\lambda(e)$ denote the length of $H(e)$. Note that $\alpha(e)$ and $\lambda(e)$ are decreasing functions of e, $e \geqq 0$.

To define $\gamma(e)$, $e > 0$, we turn to the space of coördinates (x_1, x_2, x_3) and introduce the four points

$$(e, e, 0) \qquad (-e, e, 0) \qquad (-e, -e, 0) \qquad (e, -e, 0),$$

denoting these points respectively by p_1, p_2, p_3, p_4. We form $\gamma(e)$ by joining four copies of $H(e)$ with vertices at the respective pairs of points

$$(p_1, p_2) \qquad (p_2, p_3) \qquad (p_3, p_4) \qquad (p_4, p_1)$$

and with axes having the direction cosines

$$(0, 0, 1) \qquad (0, 0, -1) \qquad (0, 0, 1) \qquad (0, 0, -1)$$

respectively. The curve $\gamma(e)$ has a continuously turning tangent, and in terms of its arc length has direction cosines satisfying a Lipschitz condition. The curve $\gamma(e)$ is accordingly admissible as a curve g.

Let T be the transformation of the (x_1, x_2, x_3)-space generated by a reflection in the plane $x_3 = 0$, followed by a rotation through 90° about the x_3-axis. Under T, $\gamma(e)$ is carried into itself. Hence the "image" under T of a minimal surface bounded by $\gamma(e)$ is a minimal surface bounded by $\gamma(e)$.

We shall consider a minimal surface S

$$x_i = x_i(u, v) \qquad (i = 1, 2, 3;\ u^2 + v^2 \leqq 1) \tag{8.5}$$

bounded by $\gamma(e)$ and shall prove the following lemma.

LEMMA 8.1. *If S is an admissible minimal surface which is bounded by $\gamma(e)$ and which is conformally invariant under T, the plane $x_3 = 0$ is tangent to a branch of S at the point $(x) = (0)$.*

Since S is conformally invariant under T, there exists a conformal mapping K of the disc $u^2 + v^2 \leqq 1$ onto itself which yields the same space image of S as does T. But K has at least one fixed point (u_0, v_0) on the unit disc, and from the nature of T, the point $(x(u_0, v_0))$ on S must be the fixed point $(x) = (0)$. The surface S may have a multiple point at the origin. To avoid difficulty we consider only that branch of S which is defined by $x_i = x_i(u, v)$ for (u, v) neighboring (u_0, v_0). The plane tangent to this branch at the origin must be in-

variant under T since S is invariant under T. But the plane $x_3 = 0$ is the only plane invariant under T, and the lemma follows directly.

The curve $\gamma(e)$ intersects the plane $x_3 = 0$ in the four points p_i, $i = 1, \cdots, 4$ in the space (x). Let q_i denote the point (u, v) corresponding to p_i under the mapping $x_i = x_i(u, v)$ of S on the unit disc. Under the hypothesis of the preceding lemma the plane $x_3 = 0$ is tangent to a branch of S at a point represented by (u_0, v_0). The "canonical arcs" on the unit disc representing the intersection of S with the plane $x_3 = 0$ must then consist of two arcs with end pairs (q_1, q_3), (q_2, q_4) and with an intersection at (u_0, v_0). In general a canonical arc with end points at the points q_iq_j will be denoted by $[q_iq_j]$. Of course $q_i \neq q_j$. The order $[q_iq_j]$ or $[q_jq_i]$ is immaterial.

More generally we shall consider the case where the minimal surface (8.5) is not necessarily conformally invariant under T. The canonical arcs on the unit disc representing the intersection of S with the plane $x_3 = 0$ will make up one of the three following sets, or sets obtained from these by a circular advance or inversion of the subscripts.

CASE I. $[q_1q_3]$, $[q_2q_4]$.

CASE II. $[q_1q_2]$, $[q_3q_4]$.

CASE III. $[q_1q_2]$, $[q_1q_3]$, $[q_1q_4]$.

We shall say that S comes under Cases I, II or III respectively, when the above canonical arcs come under these cases.

We shall prove the following lemma.

LEMMA 8.2. *Any admissible minimal surface S bounded by $\gamma(e)$ and coming under Cases* I *or* III *has an area at least* $4\alpha(0)$.

In Case I the two canonical arcs intersect in a point q. In Case III the canonical arcs intersect in q_1. If q_i and q_j are adjacent points on the circle $r = 1$ let $\{q_iq_j\}$ denote the circular arc directly between q_i and q_j. We turn first to Case III. In this case the unit disc is divided into four regions by the four closed curves.

$$[q_1q_2] + \{q_2q_1\},$$
$$[q_1q_2] + \{q_2q_3\} + [q_3q_1],$$
$$[q_1q_3] + \{q_3q_4\} + [q_4q_1],$$
$$[q_1q_4] + \{q_4q_1\}.$$

These four regions correspond under the transformation $x_i = x_i(u, v)$ to four pieces S_k of S. Each of these pieces has an area at least $\alpha(0)$. In particular the image of $[q_1q_2]$ on S is a copy H_1 of $H(e)$ with "vertices" at p_1, p_2 and an axis parallel to the positive z-axis. The orthogonal projection of S_1 onto the plane of H_1 will have an area at least $\alpha(0)$ so that S_1 has an area at least $\alpha(0)$. The same holds for S_k in general, and the Lemma follows in Case III. The

proof is similar in Case I, the canonical arcs defining a similar division of S.

The proof of the lemma is complete.

We continue with the following lemma.

LEMMA 8.3. *There exists a surface $\Sigma(e)$ bounded by $\gamma(e)$, of the topological type of the circular disc, with an area which tends to $2\alpha(0)$ as e tends to 0.*

In defining $\gamma(e)$ we joined four copies H_1, H_2, H_3, H_4 of $H(e)$. The curve $\gamma(e)$ can be spanned by a surface composed of the following three pieces. We first move H_1 orthogonal to its plane until it coincides with H_3, thereby generating the first piece. We take the second piece as the plane area bounded by H_2 and its "base," and obtain a third piece in a similar way from H_4. It is clear that these pieces can be joined so as to form a surface $\Sigma(e)$ which is the topological image of the circular disc. It is also clear that the area of $\Sigma(e)$ tends to $2\alpha(0)$ as e tends to 0.

The principal result here is the following.

THEOREM 8.1. *For all sufficiently small values of e the curve $\gamma(e)$ bounds two admissible minimal surfaces of least area defining two homotopic critical points φ_1 and φ_2 belonging to disjoint critical sets in the space Ω.*

In accordance with Lemma 8.3 there exists a positive constant e_0 such that when $|e| < e_0$ the least area of surfaces bounded by $\gamma(e)$ is less than $4\alpha(0)$. We suppose e so chosen. Let $S_1(e)$ be an admissible surface of least area. Such a surface exists in accordance with the results of Douglas and Radó. Let $S_2(e)$ be the "image" of $S_1(e)$ under the preceding space transformation T.

The surface $S_2(e)$ is minimal but it cannot be conformally equivalent to $S_1(e)$. Otherwise it would follow from Lemma 8.1 that there would be a branch of $S_1(e)$ represented by an arbitrary neighborhood of a point (u_0, v_0) for which $u_0^2 + v_0^2 < 1$, with the plane $x_3 = 0$ tangent to this branch at the point represented by (u_0, v_0). The surface $S_1(e)$ would then come under Case I and it would follow from Lemma 8.2 that the area of $S_1(e)$ would be at least $4\alpha(0)$ contrary to our choice of e. Hence $S_1(e)$ is not conformally equivalent to $S_2(e)$.

The surfaces $S_1(e)$ and $S_2(e)$ thus define distinct homotopic critical points φ_1 and φ_2 on Ω. It remains to show that φ_1 and φ_2 belong to disjoint critical sets on Ω.

If φ_1 and φ_2 belonged to the same connected critical set σ, each minimal surface $S(e)$ defined by a point φ of σ would have the area of $S_1(e)$, that is an area less than $4\alpha(0)$. Hence $S(e)$ would come under Case II by virtue of Lemma 8.2. If the canonical arcs defined by $S_1(e)$ have the form

(i) $\qquad [q_1q_2] \qquad [q_3q_4]$

the canonical arcs defined by $S_2(e)$ must have the form

(ii) $\qquad [q_2q_3] \qquad [q_4q_1]$

since $S_2(e)$ is the image of $S_1(e)$ under T. But if a surface $S(e)$ defined by φ comes under (i), minimal surfaces defined by points ψ of Ω sufficiently near φ must also come under (i). But σ is compact and connected, and it follows

that each surface $S(e)$ defined by a point of σ must come under (i). Hence φ_1 and φ_2 cannot belong to the same connected critical set and the proof of the theorem is complete.

Corollary 7.1 of §7 can then be applied to critical sets containing φ_1 and φ_2 respectively with the inference that there exists a new minimal surface of non-minimizing type, conditioned as stated in the Corollary.

THE INSTITUTE FOR ADVANCED STUDY, PRINCETON UNIVERSITY.

BIBLIOGRAPHY

1. DOUGLAS. *Solution of the problem of Plateau.* Transactions of the American Mathematical Society, 33 (1931), 263–321.
2. DOUGLAS. *Minimal surfaces of higher topological structure.* Annals of Mathematics, 40 (1939), 205–298.
3. RADÓ. *On the problem of Plateau.* Ergebnisse der Mathematik und ihrer Grenzgebiete. Volume 2 (1933), Springer, Berlin.
4. MCSHANE. *Parametrizations of saddle surfaces with application to the problem of Plateau.* Transactions of the American Mathematical Society, 35 (1933), 716–733.
5. COURANT. *Plateau's problem and Dirichlet's principle.* Annals of Mathematics, 38 (1937), 679–724.
6. SHIFFMAN. An abstract entitled: *The Plateau problem for minimal surfaces which are not relative minima.* Bulletin of the American Mathematical Society, 44 (1938), 677.
7. MORSE. *Functional topology and abstract variational theory.* Mémorial des sciences mathématiques, (1939), Gauthier-Villars, Paris.
8. MORSE. *The critical point relations in the form of group isomorphisms.* To be submitted to the Annals of Mathematics.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

A NEW PROOF OF TWO OF RAMANUJAN'S IDENTITIES[1]

BY HANS RADEMACHER AND HERBERT S. ZUCKERMAN[2]

(Received September 23, 1938)

1. When Ramanujan gave his first proof of the congruence properties

$$p(5n+4) \equiv 0 \pmod 5,$$

$$p(7n+5) \equiv 0 \pmod 7$$

of the partition function $p(n)$, he observed that these congruences would be obvious consequences of the two identities

$$p(4) + p(9)x + p(14)x^2 + \cdots = 5\,\frac{\{(1-x^5)(1-x^{10})(1-x^{15})\cdots\}^5}{\{(1-x)(1-x^2)(1-x^3)\cdots\}^6} \tag{1.11}$$

and

$$p(5) + p(12)x + p(19)x^2 + \cdots = 7\,\frac{\{(1-x^7)(1-x^{14})\cdots\}^3}{\{(1-x)(1-x^2)\cdots\}^4} + 49x\,\frac{\{(1-x^7)(1-x^{14})\cdots\}^7}{\{(1-x)(1-x^2)\cdots\}^8}. \tag{1.12}$$

However he did not give a proof of these identities. Later, making use of the theory of modular functions, Darling[3] proved the first and Mordell[4] gave proofs for both identities. Recently G. N. Watson[4a] in a comprehensive study proved these identities as well as the existence of analogous ones for all powers of 5 and 7, thereby furnishing a proof for certain further conjectures of Ramanujan.

Making use of our method, an account of which we have recently published,[5] we are able to expand the right members of (1.11) and (1.12) in power series with explicitly determined coefficients. The verification of the identities which we propose to give here will then consist merely in a direct comparison of these

---

[1] Presented to the American Mathematical Society, April 15, 1938.

[2] National Research Fellow.

[3] *Proofs of certain identities and congruences enunciated by S. Ramanujan*, Proc. London Math. Soc., (2) vol. 19 (1921), pp. 350–372.

[4] *Note on certain modular relations considered by Messrs. Ramanujan, Darling, and Rogers*, Proc. London Math. Soc., (2) vol. 20 (1922), pp. 408–416.

[4a] *Ramanujans Vermutung über Zerfällungsanzahlen*, Journal für die reine und angewandte Mathematik, vol. 179 (1938), pp. 97–128.

[5] *On the Fourier coefficients of certain modular forms of positive dimension*, Annals of Mathematics, vol. 39, no. 2 (1938), pp. 433–462.

coefficients with the corresponding coefficients[6] of the expansions of the left members. Our proof still rests on the theory of modular functions but only in so far as we make use of the transformation formula of the modular form $\eta(\tau)$.

2. The right members of (1.11) and (1.12) are made up of summands of the form

$$F_\rho(x) = \frac{\prod_{m=1}^{\infty}(1-x^{pm})^\rho}{\prod_{m=1}^{\infty}(1-x^m)^{\rho+1}}, \tag{2.1}$$

where p is one of the prime numbers 5 or 7. Since $F_\rho(x)$ can be simply expressed in terms of the function

$$f(x) = \frac{1}{\prod_{m=1}^{\infty}(1-x^m)},$$

we can make use of the identity

$$f\left(\exp\left[\frac{2\pi i h}{k} - \frac{2\pi z}{k}\right]\right) = \omega_{h,k} z^{\frac{1}{2}} \exp\left[\frac{\pi}{12kz} - \frac{\pi z}{12k}\right] f\left(\exp\left[\frac{2\pi i h'}{k} - \frac{2\pi}{kz}\right]\right), \tag{2.2}$$

with $\Re(z) > 0$, where h' is any solution of the congruence

$$hh' \equiv -1 \pmod{k}$$

and where $\omega_{h,k}$ is a certain root of unity which we shall discuss later. This identity is obtained from the theory of the transformation of the function $\eta(\tau)$ and was used by Hardy and Ramanujan[7] in their discussion of $p(n)$.

In order to apply (2.2) to $f(x^p)$ we must distinguish two cases:

A) In the case $p \nmid k$ we have to replace h, in (2.2), by ph and z by pz. We then have

$$f\left(\exp\left[\frac{2\pi i p h}{k} - \frac{2\pi p z}{k}\right]\right) = \omega_{ph,k}(pz)^{\frac{1}{2}} \exp\left[\frac{\pi}{12kpz} - \frac{\pi pz}{12k}\right] f\left(\exp\left[\frac{2\pi i h''}{k} - \frac{2\pi}{kpz}\right]\right) \tag{2.31}$$

with $phh'' \equiv -1 \pmod{k}$ or

$$ph'' \equiv h' \pmod{k}. \tag{2.32}$$

---

[6] H. Rademacher, *A convergent series for the partition function $p(n)$*, Proc. Nat. Acad. Sci., vol. 23, no. 2 (1937), pp. 78–84; *On the partition function $p(n)$*, Proc. London Math. Soc., (2) vol. 43 (1937), pp. 241–254. Cf. also pp. 455–456 of our paper cited in footnote 5.

[7] *Asymptotic formulae in combinatory analysis*, Proc. London Math. Soc., (2) vol. 17 (1918), pp. 75–115.

B) In the case $p|k$ we replace k by k/p and have

$$f\left(\exp\left[\frac{2\pi iph}{k}-\frac{2\pi pz}{k}\right]\right)$$

$$= \omega_{h,k/p}z^{\frac{1}{2}}\exp\left[\frac{\pi p}{12kz}-\frac{\pi pz}{12k}\right]f\left(\exp\left[\frac{2\pi ih'''p}{k}-\frac{2\pi p}{kz}\right]\right) \tag{2.41}$$

with $hh''' \equiv -1 \pmod{k/p}$ so that we may choose

$$h''' = h'. \tag{2.42}$$

The equations (2.2), (2.31), (2.41) enable us to write formulas for $F_\rho(x)$ with $x = \exp\left[\frac{2\pi ih}{k}-\frac{2\pi z}{k}\right]$. Again we have two cases:

A) If $p \nmid k$ we have

$$F_\rho\left(\exp\left[\frac{2\pi ih}{k}-\frac{2\pi z}{k}\right]\right) = \frac{\omega_{h,k}^{\rho+1}}{\omega_{ph,k}^{\rho}}z^{\frac{1}{2}}p^{-\rho/2}$$

$$\times\exp\left[\frac{\pi}{12k}(\rho+1)\left(\frac{1}{z}-z\right)-\frac{\pi\rho}{12k}\left(\frac{1}{pz}-pz\right)\right]\frac{f\left(\exp\left[\frac{2\pi ih'}{k}-\frac{2\pi}{kz}\right]\right)^{\rho+1}}{f\left(\exp\left[\frac{2\pi ih''}{k}-\frac{2\pi}{kpz}\right]\right)^{\rho}} \tag{2.51}$$

$$= \Omega_{h,k}^{(\rho)}z^{\frac{1}{2}}p^{-\rho/2}\exp\left[\frac{\pi}{12kpz}((p-1)\rho+p)-\frac{\pi z}{12k}(\rho+1-p\rho)\right]$$

$$\times\tilde{F}_\rho\left(\exp\left[\frac{2\pi ih''}{k}-\frac{2\pi}{kpz}\right]\right),$$

with the following abbreviations

$$\Omega_{h,k}^{(\rho)} = \frac{\omega_{h,k}^{\rho+1}}{\omega_{ph,k}^{\rho}} \qquad \text{for } p \nmid k, \tag{2.52}$$

$$\tilde{F}_\rho(x) = \frac{f(x^p)^{\rho+1}}{f(x)^\rho} = \frac{\prod_{m=1}^{\infty}(1-x^m)^\rho}{\prod_{m=1}^{\infty}(1-x^{pm})^{\rho+1}}. \tag{2.53}$$

It may be noted that we have eliminated h' by means of (2.32).

B) If $p|k$ we have

$$F_\rho\left(\exp\left[\frac{2\pi ih}{k}-\frac{2\pi z}{k}\right]\right) = \frac{\omega_{h,k}^{\rho+1}}{\omega_{h,k/p}^{\rho}}z^{\frac{1}{2}}$$

$$\times\exp\left[\frac{\pi}{12k}(\rho+1)\left(\frac{1}{z}-z\right)-\frac{\pi\rho p}{12k}\left(\frac{1}{z}-z\right)\right]\frac{f\left(\exp\left[\frac{2\pi ih'}{k}-\frac{2\pi}{kz}\right]\right)^{\rho+1}}{f\left(\exp\left[\frac{2\pi ih'p}{k}-\frac{2\pi p}{kz}\right]\right)^{\rho}} \tag{2.61}$$

$$= \Omega_{h,k}^{(\rho)}z^{\frac{1}{2}}\exp\left[\frac{\pi}{12k}((1-p)\rho+1)\left(\frac{1}{z}-z\right)\right]F_\rho\left(\exp\left[\frac{2\pi ih'}{k}-\frac{2\pi}{kz}\right]\right)$$

with the abbreviation

$$\Omega_{h,k}^{(\rho)} = \frac{\omega_{h,k}^{\rho+1}}{\omega_{h,k/p}^{\rho}} \qquad \text{for } p \mid k. \tag{2.62}$$

3. We now introduce the power series

$$F_\rho(x) = \sum_{n=0}^{\infty} a_n^{(\rho)} x^n \tag{3.11}$$

and

$$\tilde{F}_\rho(x) = \sum_{n=0}^{\infty} \tilde{a}_n^{(\rho)} x^n \tag{3.12}$$

for the functions (2.1) and (2.53). From the definitions we have, at once,

$$a_0^{(\rho)} = 1, \quad \tilde{a}_0^{(\rho)} = 1, \quad \tilde{a}_1^{(\rho)} = -\rho, \quad \tilde{a}_2^{(\rho)} = \tfrac{1}{2}\rho(\rho-3), \tag{3.13}$$

providing $p > 2$.

Our next purpose is to determine the coefficients $a_n^{(\rho)}$ of (3.11). This is done by our method quoted above. As all the steps are quite analogous we shall omit the details and merely display the essential stages. We have

$$a_n^{(\rho)} = \frac{1}{2\pi i}\int \frac{F_\rho(x)}{x^{n+1}}\,dx, \qquad |x| = \exp\left[-\frac{2\pi}{N^2}\right],$$

$$= \sum_{h,k\le N}{}' \frac{1}{2\pi i}\int_{\xi_{h,k}} \frac{F_\rho(x)}{x^{n+1}}\,dx,$$

where $\xi_{h,k}$ is the Farey arc corresponding to the value h/k. We designate the end-points of this arc, measured by the angle φ from $2\pi h/k$, as $-\vartheta'_{h,k}$ and $\vartheta''_{h,k}$. The symbol $\sum'_{h,k\le N}$ designates the sum over all h and k such that $0 \le h < k \le N$ and $(h, k) = 1$.

Writing $w = N^{-2} - i\varphi$ we have

$$a_n^{(\rho)} = \sum_{h,k\le N}{}' e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} F_\rho\left(\exp\left[\frac{2\pi i h}{k} - 2\pi w\right]\right) e^{2\pi n w}\,d\varphi.$$

In the sum over k we now apply (2.51) or (2.61) according as $p \nmid k$ or $p|k$ and have, using the notation of (3.11) and (3.12),

$$a_n^{(\rho)} = p^{-\rho/2} \sum_{\substack{k\le N\\ p\nmid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[\frac{2\pi}{k^2 p w}\,\frac{(p-1)\rho + p}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} \tilde{a}_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h''}{k} - \frac{2\pi}{k^2 p w}\right)\right] d\varphi$$

$$+ \sum_{\substack{k\le N\\ p\mid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[-\frac{2\pi}{k^2 w}\,\frac{\rho(p-1)-1}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} a_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h'}{k} - \frac{2\pi}{k^2 w}\right)\right] d\varphi.$$

We can now follow exactly the reasoning of sections 3, 4, and 5 of our previous paper. The factor $w^{\frac{1}{2}}$ above shows that, corresponding to the dimension r, we are to take the positive value $\frac{1}{2}$. From section 3 we see that only terms in which a factor $e^{A/w}$ with $A > 0$ appears will furnish a contribution to the final result when we let N tend to infinity. This means, in particular, that the terms corresponding to $p|k$ will drop out. The condition (4.1) of that paper becomes $(\rho(p-1)-1)/24 + n > 0$ and hence is satisfied for $n \geqq 0$. Thus, in analogy to (5.5), we find the result

$$(3.21)\qquad a_n^{(\rho)} = \frac{2\pi}{p^{\rho/2}} \sum_{m=0}^{\mu} \tilde{a}_m^{(\rho)} \sum_{\substack{k=1 \\ p\nmid k}}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i}{k}(hn-h'm)} \left(\frac{\frac{(p-1)\rho+p}{24p} - \frac{m}{p}}{\frac{\rho(p-1)-1}{24} + n} \right)^{\frac{1}{2}}$$
$$\times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{(p-1)\rho+p-24m}{24p}\left(\frac{\rho(p-1)-1}{24}+n\right)} \right),$$

where μ is the greatest integer m for which

$$(3.22)\qquad (p-1)\rho + p - 24m > 0.$$

The values of the coefficients of the left members of (1.11) and (1.12), with which (3.21) is to be compared, are[8]

$$(3.3)\qquad p(pn+\lambda) = 2\pi \sum_{k=1}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(pn+\lambda)} \left(\frac{\frac{1}{24}}{pn+\lambda-\frac{1}{24}} \right)^{\frac{1}{2}}$$
$$\times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{1}{24}\left(pn+\lambda-\frac{1}{24}\right)} \right).$$

4. At this point we separate the two cases $p = 5$ and $p = 7$. The first identity, (1.11), can be written as

$$(4.11)\qquad \sum_{n=0}^{\infty} p(5n+4)\, x^n = 5\, F_5(x) = 5 \sum_{n=0}^{\infty} a_n^{(5)} x^n,$$

and hence will be proved if we show

$$(4.12)\qquad p(5n+4) = 5a_n^{(5)}.$$

In this case we have

$$(4.13)\qquad p = 5, \qquad \lambda = 4, \qquad \rho = 5,$$

and hence, by (3.22),

$$(4.14)\qquad \mu = 1.$$

[8] We use the form in which the series for $p(n)$ is given on pp. 455–456 of our paper quoted in footnote 5.

with the abbreviation

$$\Omega_{h,k}^{(\rho)} = \frac{\omega_{h,k}^{\rho+1}}{\omega_{h,k/p}^{\rho}} \qquad \text{for } p \mid k. \tag{2.62}$$

3. We now introduce the power series

$$F_\rho(x) = \sum_{n=0}^{\infty} a_n^{(\rho)} x^n \tag{3.11}$$

and

$$\tilde{F}_\rho(x) = \sum_{n=0}^{\infty} \tilde{a}_n^{(\rho)} x^n \tag{3.12}$$

for the functions (2.1) and (2.53). From the definitions we have, at once,

$$a_0^{(\rho)} = 1, \quad \tilde{a}_0^{(\rho)} = 1, \quad \tilde{a}_1^{(\rho)} = -\rho, \quad \tilde{a}_2^{(\rho)} = \tfrac{1}{2}\rho(\rho-3), \tag{3.13}$$

providing $p > 2$.

Our next purpose is to determine the coefficients $a_n^{(\rho)}$ of (3.11). This is done by our method quoted above. As all the steps are quite analogous we shall omit the details and merely display the essential stages. We have

$$a_n^{(\rho)} = \frac{1}{2\pi i}\int \frac{F_\rho(x)}{x^{n+1}}\,dx, \qquad |x| = \exp\left[-\frac{2\pi}{N^2}\right],$$

$$= \sum_{h,k\le N}{}' \frac{1}{2\pi i}\int_{\xi_{h,k}} \frac{F_\rho(x)}{x^{n+1}}\,dx,$$

where $\xi_{h,k}$ is the Farey arc corresponding to the value h/k. We designate the end-points of this arc, measured by the angle φ from $2\pi h/k$, as $-\vartheta'_{h,k}$ and $\vartheta''_{h,k}$. The symbol $\sum'_{h,k\le N}$ designates the sum over all h and k such that $0 \le h < k \le N$ and $(h, k) = 1$.

Writing $w = N^{-2} - i\varphi$ we have

$$a_n^{(\rho)} = \sum_{h,k\le N}{}' e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} F_\rho\left(\exp\left[\frac{2\pi i h}{k} - 2\pi w\right]\right) e^{2\pi n w}\,d\varphi.$$

In the sum over k we now apply (2.51) or (2.61) according as $p \nmid k$ or $p \mid k$ and have, using the notation of (3.11) and (3.12),

$$a_n^{(\rho)} = p^{-\rho/2} \sum_{\substack{k\le N\\ p\nmid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[\frac{2\pi}{k^2 p w}\,\frac{(p-1)\rho + p}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} \tilde{a}_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h''}{k} - \frac{2\pi}{k^2 p w}\right)\right] d\varphi$$

$$+ \sum_{\substack{k\le N\\ p\mid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[-\frac{2\pi}{k^2 w}\,\frac{\rho(p-1)-1}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} a_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h'}{k} - \frac{2\pi}{k^2 w}\right)\right] d\varphi.$$

We can now follow exactly the reasoning of sections 3, 4, and 5 of our previous paper. The factor $w^{\frac{1}{2}}$ above shows that, corresponding to the dimension r, we are to take the positive value $\frac{1}{2}$. From section 3 we see that only terms in which a factor $e^{A/w}$ with $A > 0$ appears will furnish a contribution to the final result when we let N tend to infinity. This means, in particular, that the terms corresponding to $p|k$ will drop out. The condition (4.1) of that paper becomes $(\rho(p-1)-1)/24 + n > 0$ and hence is satisfied for $n \geqq 0$. Thus, in analogy to (5.5), we find the result

$$a_n^{(\rho)} = \frac{2\pi}{p^{\rho/2}} \sum_{m=0}^{\mu} \bar{a}_m^{(\rho)} \sum_{\substack{k=1 \\ p\nmid k}}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i}{k}(hn-h''m)} \left(\frac{\frac{(p-1)\rho+p}{24p} - \frac{m}{p}}{\frac{\rho(p-1)-1}{24} + n} \right)^{\frac{1}{4}} \tag{3.21}$$

$$\times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{(p-1)\rho+p-24m}{24p}\left(\frac{\rho(p-1)-1}{24}+n\right)} \right),$$

where μ is the greatest integer m for which

$$(p-1)\rho + p - 24m > 0. \tag{3.22}$$

The values of the coefficients of the left members of (1.11) and (1.12), with which (3.21) is to be compared, are[8]

$$p(pn+\lambda) = 2\pi \sum_{k=1}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(pn+\lambda)} \left(\frac{\frac{1}{24}}{pn+\lambda-\frac{1}{24}} \right)^{\frac{1}{4}} \tag{3.3}$$

$$\times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{1}{24}\left(pn+\lambda-\frac{1}{24}\right)} \right).$$

4. At this point we separate the two cases $p = 5$ and $p = 7$. The first identity, (1.11), can be written as

$$\sum_{n=0}^{\infty} p(5n+4)\, x^n = 5\, F_5(x) = 5 \sum_{n=0}^{\infty} a_n^{(5)} x^n, \tag{4.11}$$

and hence will be proved if we show

$$p(5n+4) = 5a_n^{(5)}. \tag{4.12}$$

In this case we have

$$p = 5, \qquad \lambda = 4, \qquad \rho = 5, \tag{4.13}$$

and hence, by (3.22),

$$\mu = 1. \tag{4.14}$$

[8] We use the form in which the series for $p(n)$ is given on pp. 455–456 of our paper quoted in footnote 5.

with the abbreviation

$$\Omega_{h,k}^{(\rho)} = \frac{\omega_{h,k}^{\rho+1}}{\omega_{h,k/p}^{\rho}} \qquad \text{for } p \mid k. \tag{2.62}$$

3. We now introduce the power series

$$F_\rho(x) = \sum_{n=0}^{\infty} a_n^{(\rho)} x^n \tag{3.11}$$

and

$$\tilde{F}_\rho(x) = \sum_{n=0}^{\infty} \tilde{a}_n^{(\rho)} x^n \tag{3.12}$$

for the functions (2.1) and (2.53). From the definitions we have, at once,

$$a_0^{(\rho)} = 1, \quad \tilde{a}_0^{(\rho)} = 1, \quad \tilde{a}_1^{(\rho)} = -\rho, \quad \tilde{a}_2^{(\rho)} = \tfrac{1}{2}\rho(\rho-3), \tag{3.13}$$

providing $p > 2$.

Our next purpose is to determine the coefficients $a_n^{(\rho)}$ of (3.11). This is done by our method quoted above. As all the steps are quite analogous we shall omit the details and merely display the essential stages. We have

$$a_n^{(\rho)} = \frac{1}{2\pi i}\int \frac{F_\rho(x)}{x^{n+1}}\,dx, \qquad |x| = \exp\left[-\frac{2\pi}{N^2}\right],$$

$$= \sum_{h,k\le N}{}' \frac{1}{2\pi i}\int_{\xi_{h,k}} \frac{F_\rho(x)}{x^{n+1}}\,dx,$$

where $\xi_{h,k}$ is the Farey arc corresponding to the value h/k. We designate the end-points of this arc, measured by the angle φ from $2\pi h/k$, as $-\vartheta'_{h,k}$ and $\vartheta''_{h,k}$. The symbol $\sum'_{h,k\le N}$ designates the sum over all h and k such that $0 \le h < k \le N$ and $(h, k) = 1$.

Writing $w = N^{-2} - i\varphi$ we have

$$a_n^{(\rho)} = \sum_{h,k\le N}{}' e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} F_\rho\left(\exp\left[\frac{2\pi i h}{k} - 2\pi w\right]\right) e^{2\pi n w}\,d\varphi.$$

In the sum over k we now apply (2.51) or (2.61) according as $p \nmid k$ or $p \mid k$ and have, using the notation of (3.11) and (3.12),

$$a_n^{(\rho)} = p^{-\rho/2} \sum_{\substack{k\le N\\ p\nmid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[\frac{2\pi}{k^2 p w}\,\frac{(p-1)\rho + p}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} \tilde{a}_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h''}{k} - \frac{2\pi}{k^2 p w}\right)\right] d\varphi$$

$$+ \sum_{\substack{k\le N\\ p\mid k}} k^{\frac{1}{2}} \sum_{h \bmod k}{}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i h n}{k}} \int_{-\vartheta'_{h,k}}^{\vartheta''_{h,k}} w^{\frac{1}{2}} \exp\left[-\frac{2\pi}{k^2 w}\,\frac{\rho(p-1)-1}{24}\right.$$

$$\left. + 2\pi w\left(\frac{\rho(p-1)-1}{24} + n\right)\right] \sum_{m=0}^{\infty} a_m^{(\rho)} \exp\left[m\left(\frac{2\pi i h'}{k} - \frac{2\pi}{k^2 w}\right)\right] d\varphi.$$

We can now follow exactly the reasoning of sections 3, 4, and 5 of our previous paper. The factor $w^{\frac{1}{2}}$ above shows that, corresponding to the dimension r, we are to take the positive value $\frac{1}{2}$. From section 3 we see that only terms in which a factor $e^{A/w}$ with $A > 0$ appears will furnish a contribution to the final result when we let N tend to infinity. This means, in particular, that the terms corresponding to $p|k$ will drop out. The condition (4.1) of that paper becomes $(\rho(p-1)-1)/24 + n > 0$ and hence is satisfied for $n \geqq 0$. Thus, in analogy to (5.5), we find the result

$$a_n^{(\rho)} = \frac{2\pi}{p^{\rho/2}} \sum_{m=0}^{\mu} \tilde{a}_m^{(\rho)} \sum_{\substack{k=1 \\ p\nmid k}}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \Omega_{h,k}^{(\rho)} e^{-\frac{2\pi i}{k}(hn-h'm)} \left(\frac{\frac{(p-1)\rho+p}{24p} - \frac{m}{p}}{\frac{\rho(p-1)-1}{24} + n} \right)^{\frac{1}{4}} \times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{(p-1)\rho+p-24m}{24p}\left(\frac{\rho(p-1)-1}{24}+n\right)} \right), \tag{3.21}$$

where μ is the greatest integer m for which

$$(p-1)\rho + p - 24m > 0. \tag{3.22}$$

The values of the coefficients of the left members of (1.11) and (1.12), with which (3.21) is to be compared, are[8]

$$p(pn+\lambda) = 2\pi \sum_{k=1}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(pn+\lambda)} \left(\frac{\frac{1}{24}}{pn+\lambda-\frac{1}{24}} \right)^{\frac{1}{4}} \times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{1}{24}\left(pn+\lambda-\frac{1}{24}\right)} \right). \tag{3.3}$$

4. At this point we separate the two cases $p = 5$ and $p = 7$. The first identity, (1.11), can be written as

$$\sum_{n=0}^{\infty} p(5n+4)\, x^n = 5\, F_5(x) = 5 \sum_{n=0}^{\infty} a_n^{(5)} x^n, \tag{4.11}$$

and hence will be proved if we show

$$p(5n+4) = 5a_n^{(5)}. \tag{4.12}$$

In this case we have

$$p = 5, \qquad \lambda = 4, \qquad \rho = 5, \tag{4.13}$$

and hence, by (3.22),

$$\mu = 1. \tag{4.14}$$

[8] We use the form in which the series for $p(n)$ is given on pp. 455–456 of our paper quoted in footnote 5.

Also from (3.13) we have

$$\tilde{a}_0^{(5)} = 1, \qquad \tilde{a}_1^{(5)} = -5. \tag{4.15}$$

Setting these values in (3.21) we have

$$\begin{aligned} 5a_n^{(5)} &= \frac{2\pi}{5^{3/2}} \sum_{\substack{k=1 \\ 5\nmid k}}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i h n}{k}} \left(\frac{\frac{5}{24}}{n + \frac{19}{24}} \right)^{\frac{3}{4}} I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{5}{24}\left(n + \frac{19}{24}\right)} \right) \\ &\quad - \frac{2\pi}{5^{\frac{1}{2}}} \sum_{\substack{k=1 \\ 5\nmid k}}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i}{k}(hn - h'')} \left(\frac{\frac{1}{120}}{n + \frac{19}{24}} \right)^{\frac{3}{4}} I_{3/2}\left(\frac{4\pi}{5k} \sqrt{\frac{5}{24}\left(n + \frac{19}{24}\right)} \right) \end{aligned} \tag{4.21}$$

which is to be compared with

$$\begin{aligned} p(5n+4) &= 2\pi \sum_{k=1}^{\infty} \frac{1}{k} \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(5n+4)} \left(\frac{\frac{1}{24}}{5\left(n + \frac{19}{24}\right)} \right)^{\frac{3}{4}} \\ &\quad \times I_{3/2}\left(\frac{4\pi}{k} \sqrt{\frac{5}{24}\left(n + \frac{19}{24}\right)} \right), \end{aligned} \tag{4.22}$$

which we have by (3.3).

Comparing terms in (4.21) and (4.22) which have the same arguments in the Bessel functions, we see that our proof of (4.12), and hence of (1.11), will be accomplished as soon as we have shown the following three facts:

I. $$\sum_{h \bmod k}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i h n}{k}} = \sum_{h \bmod k} \omega_{h,k} e^{-\frac{2\pi i h}{k}(5n+4)}, \qquad 5 \nmid k, \qquad p = 5;$$

II. $$-\frac{1}{5^{\frac{1}{2}}} \sum_{h \bmod k}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i}{k}(hn - h'')} = \frac{1}{5} \sum_{h \bmod 5k}' \omega_{h,5k} e^{-\frac{2\pi i h}{5k}(5n+4)}, \qquad 5 \nmid k, \qquad p = 5;$$

III. $$\sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(5n+4)} = 0 \qquad \text{for} \qquad 5^2 \mid k.$$

It is to be remembered that $\Omega_{h,k}^{(\rho)}$ depends on the value of p and that these equalities are asserted only for $p = 5$.

The second identity (1.12) is treated in a similar manner. We write it as

$$\sum_{n=0}^{\infty} p(7n+5)x^n = 7F_3(x) + 49xF_7(x) = 7a_0^{(3)} + \sum_{n=1}^{\infty} (7a_n^{(3)} + 49a_{n-1}^{(7)})x^n. \tag{4.31}$$

By (3.13) we have $a_0^{(3)} = 1$ and hence $p(5) = 7 = 7a_0^{(3)}$ and we need only show

$$p(7n+5) = 7a_n^{(3)} + 49a_{n-1}^{(7)}, \qquad n \geqq 1. \tag{4.32}$$

For $F_3(x)$ we have

$$p = 7, \quad \rho = 3, \quad \mu = 1, \quad \tilde{a}_0^{(3)} = 1, \quad \tilde{a}_1^{(3)} = -3,$$

and hence, by (3.21),

$$7a_n^{(3)} = \frac{2\pi}{7^{\frac{1}{2}}}\sum_{\substack{k=1\\7\nmid k}}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}hn}\left(\frac{\frac{25}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}} I_{3/2}\left(\frac{20\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right)$$

(4.41)

$$-\frac{6\pi}{7^{\frac{1}{2}}}\sum_{\substack{k=1\\7\nmid k}}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')}\left(\frac{\frac{1}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}} I_{3/2}\left(\frac{4\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right).$$

For $F_7(x)$ we have

$$p = 7, \quad \rho = 7, \quad \mu = 2, \quad \tilde{a}_0^{(7)} = 1, \quad \tilde{a}_1^{(7)} = -7, \quad \tilde{a}_2^{(7)} = 14,$$

and then we find

$$49a_{n-1}^{(7)} = \frac{2\pi}{7^{3/2}}\sum_{\substack{k=1\\7\nmid k}}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}h(n-1)}\left(\frac{\frac{49}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}}$$
$$\times I_{3/2}\left(\frac{28\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right)$$

$$-\frac{2\pi}{7^{\frac{1}{2}}}\sum_{\substack{k=1\\7\nmid k}}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}(h(n-1)-h'')}\left(\frac{\frac{25}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}} \tag{4.42}$$
$$\times I_{3/2}\left(\frac{20\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right)$$

$$+\frac{4\pi}{7^{\frac{1}{2}}}\sum_{\substack{k=1\\7\nmid k}}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}(h(n-1)-2h'')}\left(\frac{\frac{1}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}}$$
$$\times I_{3/2}\left(\frac{4\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right).$$

On the other side of (4.32) we have

$$p(7n+5) = 2\pi\sum_{k=1}^{\infty}\frac{1}{k}\sum_{h \bmod k}{}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(7n+5)}\left(\frac{\frac{1}{24}}{7\left(n+\frac{17}{24}\right)}\right)^{\frac{1}{4}} \tag{4.43}$$
$$\times I_{3/2}\left(\frac{28\pi}{k}\sqrt{\frac{1}{7\cdot 24}\left(n+\frac{17}{24}\right)}\right).$$

If we now set (4.41), (4.42) and (4.43) in (4.32) and again compare terms having the same arguments in the Bessel functions we see that our proof of (4.32), and hence of (1.12), is now reduced to show the four facts:

$$\text{I}' \qquad \sum_{h \bmod k}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}h(n-1)} = \sum_{k \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(7n+5)}, \quad 7 \nmid k, p = 7;$$

$$\text{II}' \qquad -\frac{3}{7^{\frac{1}{2}}} \sum_{h \bmod k}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} + \frac{2}{7^{\frac{1}{2}}} \sum_{h \bmod k}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}(h(n-1)-2h'')}$$
$$= \frac{1}{7} \sum_{h \bmod 7k}' \omega_{h,7k} e^{-\frac{2\pi i h}{7k}(7n+5)}, \qquad 7 \nmid k, \qquad p = 7;$$

$$\text{III}' \qquad \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(7n+5)} = 0 \quad \text{for} \quad 7^2 | k;$$

$$\text{IV}' \qquad \sum_{h \bmod k}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}hn} - \sum_{h \bmod k}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i}{k}(h(n-1)-h'')} = 0, \quad 7 \nmid k, \quad p = 7.$$

5. For the proof of the relations stated in section 4 we shall need the two following lemmas. These lemmas are based on some recent results of D. H. Lehmer.[9] It is convenient to use the notation

$$A_k(n) = \sum_{h \bmod k}' \omega_{h,k} e^{-\frac{2\pi i h}{k}n} \tag{5.1}$$

which was introduced by Hardy and Ramanujan[10] and which is in agreement with the notation of Lehmer's paper.

LEMMA 1: *Let p be a prime, $p > 3$, $p \nmid k$ and let*

$$24\lambda \equiv 1 \pmod{p}. \tag{5.21}$$

Then

$$A_{pk}(pn + \lambda) = A_p(\lambda) A_k(n_1) \tag{5.22}$$

where n_1 is any solution of the congruence

$$p^2 n_1 \equiv pn + (p^2 - 1 + 24\lambda)/24 \pmod{k}. \tag{5.23}$$

It is to be noted that the coefficients in (5.23) *are integers and have the common factor p.*

PROOF: It is obvious from (5.1) that $A_k(n)$ is a periodic function of n with the period k. Hence in the course of the proof we need determine the value of the argument only modulo the subscript. We have, with Lehmer, to distinguish three cases according to the divisibility of k by powers of 2.

---

[9] *On the series for the partition function*, Transactions of the American Mathematical Society, vol. 43, no. 2 (1938), pp. 271–295.

[10] Loc. cit. footnote 7, especially p. 85.

A) If k is odd we have, by Lehmer's Theorem 1,

$$A_{pk}(pn + \lambda) = A_k(n_1)A_p(n_2) \tag{5.31}$$

with

$$pn + \lambda \equiv p^2n_1 + k^2n_2 - (p^2 + k^2 - 1)/24 \pmod{pk}. \tag{5.32}$$

Hence we have

$$24\lambda \equiv 24k^2n_2 - k^2 + 1 \pmod{p},$$

$$24\lambda - 1 \equiv k^2(24n_2 - 1) \pmod{p}$$

and then, by (5.21),

$$\begin{aligned} 24n_2 - 1 &\equiv 0 \pmod{p} \\ n_2 &\equiv \lambda \pmod{p}. \end{aligned} \tag{5.33}$$

From (5.32) we also have

$$pn + \lambda \equiv p^2n_1 - (p^2 + k^2 - 1)/24 \pmod{k}. \tag{5.34}$$

If $3 \nmid k$ this yields

$$24(pn + \lambda) \equiv 24p^2n_1 - p^2 + 1 \pmod{k},$$

$$pn + \lambda \equiv p^2n_1 - (p^2 - 1)/24 \pmod{k} \tag{5.35}$$

while if $3 \mid k$ we have

$$8(pn + \lambda) \equiv 8p^2n_1 - (p^2 - 1)/3 \pmod{k}$$

$$pn + \lambda \equiv p^2n_1 - (p^2 - 1)/24 \pmod{k}. \tag{5.36}$$

The formulas (5.33), (5.35) and (5.36) complete the proof for k odd.

B) If $k = 2^\alpha k_1$, $\alpha > 1$, k_1 odd, we use Lehmer's Theorems 1 and 2 and find the following relations:

$$A_{pk}(pn + \lambda) = (-1)^{2^{\alpha-2}}A_{2^\alpha}(n_2)A_{pk_1}(n_3) \tag{5.41}$$

with

$$pn + \lambda \equiv p^2k_1^2n_2 + 2^{2\alpha}n_3 - (p^2k_1^2 - 1 + 2^{2\alpha})/24 \pmod{2^\alpha pk_1}, \tag{5.42}$$

$$A_{pk_1}(n_3) = A_p(n_4)A_{k_1}(n_5) \tag{5.43}$$

with

$$n_3 \equiv p^2n_5 + k_1^2n_4 - (p^2 + k_1^2 - 1)/24 \pmod{pk_1}, \tag{5.44}$$

and

$$A_k(n_1) = A_{2^\alpha k_1}(n_1) = (-1)^{2^{\alpha-2}}A_{k_1}(n_5)A_{2^\alpha}(n_2) \tag{5.45}$$

with

$$n_1 \equiv k_1^2n_2 + 2^{2\alpha}n_5 - (k_1^2 - 1 + 2^{2\alpha})/24 \pmod{2^\alpha k_1}. \tag{5.46}$$

From (5.41), (5.43) and (5.45) we have

$$A_{pk}(pn + \lambda) = A_p(n_4)A_k(n_1) \tag{5.47}$$

while (5.42) and (5.44) yield

$$24\lambda - 1 \equiv 2^{2\alpha}(24n_3 - 1) \pmod{p}$$

and

$$24n_3 - 1 \equiv k_1^2(24n_4 - 1) \pmod{p}$$

and hence

$$2^{2\alpha}k_1^2(24n_4 - 1) \equiv 24\lambda - 1 \equiv 0 \pmod{p}$$

or

$$n_4 \equiv \lambda \pmod{p} \tag{5.48}$$

in virtue of (5.21). From (5.42) and (5.46) we get

$$pn + \lambda \equiv p^2k_1^2n_2 - (p^2k_1^2 - 1 + 2^{2\alpha})/24 \pmod{2^\alpha}$$

and

$$n_1 \equiv k_1^2n_2 - (k_1^2 - 1 + 2^{2\alpha})/24 \pmod{2^\alpha}$$

which, on elimination of n_2, yields

$$\begin{aligned} p^2n_1 &\equiv pn + \lambda + (1 - 2^{2\alpha})(p^2 - 1)/24 \\ &\equiv pn + \lambda + (p^2 - 1)/24 \pmod{2^\alpha}. \end{aligned} \tag{5.49}$$

Finally we use (5.42), (5.44) and (5.46) to obtain

$$\begin{aligned} pn + \lambda &\equiv 2^{2\alpha}n_3 - (2^{2\alpha} - 1)/24 \pmod{k_1}, \\ n_3 &\equiv p^2n_5 - (p^2 - 1)/24 \pmod{k_1}, \\ n_1 &\equiv 2^{2\alpha}n_5 - (2^{2\alpha} - 1)/24 \pmod{k_1}. \end{aligned}$$

Eliminating n_3 and n_5 we get

$$\begin{aligned} pn + \lambda &\equiv 2^{2\alpha}p^2n_5 - (2^{2\alpha}p^2 - 1)/24 \pmod{k_1} \\ &\equiv p^2n_1 - (p^2 - 1)/24 \pmod{k_1} \end{aligned}$$

or

$$p^2n_1 \equiv pn + \lambda + (p^2 - 1)/24 \pmod{k_1}$$

which, together with (5.49) yields (5.23). This with (5.47) and (5.48) completes the proof for this case.

C) If $k = 2k_1$, k_1 odd, we use Lehmer's Theorems 1 and 4 to obtain

$$A_{pk}(pn + \lambda) = A_{2pk_1}(pn + \lambda) = A_2(n_2)A_{pk_1}(n_3)$$

with

$$pn + \lambda \equiv 4n_3 + pk_1n_2 + (p^2k_1^2 - 1)/8$$
$$\equiv 4n_3 + p^2k_1^2n_2 + (p^2k_1^2 - 1)/8 \pmod{2pk_1},$$
$$A_{pk_1}(n_3) = A_p(n_4)A_{k_1}(n_5)$$

with

$$n_3 \equiv p^2n_5 + k_1^2n_4 - (p^2 + k_1^2 - 1)/24 \pmod{pk_1},$$

and

$$A_k(n_1) = A_{2k_1}(n_1) = A_2(n_2)A_{k_1}(n_5)$$

with

$$n_1 \equiv 4n_5 + k_1n_2 + (k_1^2 - 1)/8 \equiv 4n_5 + k_1^2n_2 + (k_1^2 - 1)/8 \pmod{2k_1}.$$

These formulas can be combined in a manner similar to that used in case B thus completing the proof of the lemma.

LEMMA 2: *Under the conditions of Lemma* 1 *and with* $\beta > 1$ *we have*

$$A_{p^\beta k}(pn + \lambda) = 0. \tag{5.51}$$

PROOF: We again treat three cases as in the proof of Lemma 1.

A) If k is odd we use Lehmer's Theorem 1 to obtain

$$A_{p^\beta k}(pn + \lambda) = A_k(n_1)A_{p^\beta}(n_2)$$

with

$$pn + \lambda \equiv p^{2\beta}n_1 + k^2n_2 - (p^{2\beta} + k^2 - 1)/24 \pmod{p^\beta k},$$

from which we get

$$\lambda \equiv k^2n_2 - (k^2 - 1)/24 \pmod{p},$$
$$24\lambda - 1 \equiv k^2(24n_2 - 1) \pmod{p},$$
$$24n_2 - 1 \equiv 0 \pmod{p}. \tag{5.52}$$

Now, however, (5.52) implies

$$A_{p^\beta}(n_2) = 0$$

by Lehmer's Theorem 5 and hence our proof is complete for this case.

B) If $k = 2^\alpha k_1$, $\alpha > 1$, k_1 odd, we have, using Lehmer's Theorem 2,

$$A_{p^\beta k}(pn + \lambda) = A_{2^\alpha p^\beta k_1}(pn + \lambda) = (-1)^{2^{\alpha-2}}A_{2^\alpha}(n_2)A_{p^\beta k_1}(n_3)$$

with

$$pn + \lambda \equiv p^{2\beta}k_1^2n_2 + 2^{2\alpha}n_3 - (p^{2\beta}k_1^2 + 2^{2\alpha} - 1)/24 \pmod{2^\alpha p^\beta k_1}$$

and hence

$$24\lambda - 1 \equiv 2^{2\alpha}(24n_3 - 1) \pmod{p},$$

$$24n_3 - 1 \equiv 0 \pmod{p},$$

$$n_3 \equiv \lambda \pmod{p}.$$

Therefore we have

$$A_{p^\beta k_1}(n_3) = 0$$

by case A which we have just proved.

C) If $k = 2k_1$, k_1 odd, Lehmer's Theorem 4 can be used to reduce this case to case A just as before.

6. We turn now to the proof of the relations asserted in section 4. We first consider those having to do with the modulus 5.

We have[11] for odd k:

$$\omega_{h,k} = \left(\frac{-h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k - \frac{1}{k}\right)(2h - h' + h^2h')\right)\right\} \tag{6.11}$$

and for odd h:

$$\omega_{h,k} = \left(\frac{-k}{h}\right) \exp\left\{-\pi i\left(\frac{1}{4}(2 - hk - h) + \frac{1}{12}\left(k - \frac{1}{k}\right)(2h - h' + h^2h')\right)\right\}, \tag{6.12}$$

and therefore by (2.52) for odd k, $5 \nmid k$, we have

$$\Omega_{h,k}^{(5)} = \frac{\omega_{h,k}^6}{\omega_{5h,k}^5} = \left(\frac{-5h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k - \frac{1}{k}\right)(12h - 6h' + 6h^2h' - 50h + 5h'' - 125h^2h'')\right)\right\}$$

with

$$5hh'' \equiv -1 \pmod{k}. \tag{6.21}$$

From this congruence we have

$$h' \equiv 5h'' \pmod{k} \tag{6.22}$$

and hence we may write

$$\Omega_{h,k}^{(5)} = \left(\frac{-5h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k - \frac{1}{k}\right)(-38h - 25h'' - 95h^2h'')\right)\right\}. \tag{6.23}$$

[11] Loc. cit. footnote 7, especially p. 85.

If we now replace h by $5h$ we see from the definition (6.21) that we must replace $5h''$ by h'' at the same time. Doing this we get, from (6.23),

$$\Omega^{(5)}_{5h,k} = \left(\frac{-h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k-\frac{1}{k}\right)(-190h - 5h'' - 475h^2h'')\right)\right\}$$

$$= \left(\frac{-h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1)\right.\right.$$

$$\left.\left. + \frac{1}{12}\left(k-\frac{1}{k}\right)(-94h - h' + h^2h' - 96h(hh'+1))\right)\right\}$$

where we have made use of (6.22). Since, by definition, $hh' + 1 \equiv 0 \pmod{k}$, this equation goes over into the form

$$\Omega^{(5)}_{5h,k} = \left(\frac{-h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k-\frac{1}{k}\right)(2h - h' + h^2h')\right)\right\} e^{-\frac{\pi i 8h}{k}}$$

$$= \omega_{h,k}\, e^{-\frac{\pi i 8h}{k}}.$$

Now $5h$ runs, with h, through a reduced set of residues modulo k since $5 \nmid k$ and hence we have

$$\sum'_{h \bmod k} \Omega^{(5)}_{h,k} e^{-\frac{2\pi i h n}{k}} = \sum'_{h \bmod k} \Omega^{(5)}_{5h,k} e^{-\frac{2\pi i 5h n}{k}}$$

$$= \sum'_{h \bmod k} \omega_{h,k}\, e^{-\frac{2\pi i h}{k}(5n+4)}$$

which proves I for odd k.

For the case k even we use (6.12) and find, as before,

$$\Omega^{(5)}_{h,k} = \left(\frac{-k}{5h}\right) \exp\left\{-\pi i\left(\frac{1}{4}(2 + 19hk + 19h)\right.\right.$$

(6.3)

$$\left.\left. + \frac{1}{12}\left(k-\frac{1}{k}\right)(-38h - 25h'' - 95h^2h'')\right)\right\},$$

$$\Omega^{(5)}_{5h,k} = \left(\frac{-k}{h}\right) \exp\left\{-\pi i\left(\frac{1}{4}(2 + 95hk + 95h)\right.\right.$$

$$\left.\left. + \frac{1}{12}\left(k-\frac{1}{k}\right)(-190h - 5h'' - 475h^2h'')\right)\right\}$$

$$= \left(\frac{-k}{h}\right) \exp\left\{-\pi i\left(\frac{1}{4}(2 - hk - h)\right.\right.$$

$$\left.\left. + \frac{1}{12}\left(k-\frac{1}{k}\right)(2h - h' + h^2h')\right)\right\} e^{-\frac{\pi i 8h}{k}}$$

$$= \omega_{h,k}\, e^{-\frac{\pi i 8h}{k}}.$$

The remainder of the proof is the same as in the previous case. Thus I is proved also for even k.

7. For the proof of II we first use Lemma 1. With $p = 5$ we have

$$\lambda = 4,$$
$$5^2 n_1 \equiv 5n + 5 \pmod{k},$$
$$n_1 \equiv 5^*(n+1) \pmod{k},$$

where 5^* is the reciprocal of 5 modulo k, that is

$$5 \cdot 5^* \equiv 1 \pmod{k}.$$

Then for $5 \nmid k$ we have

$$A_{5k}(5n+4) = A_5(4) A_k(5^*(n+1)).$$

From Lehmer's Theorem 5 we find the value

$$A_5(4) = \left(\frac{3}{5}\right) 5^{\frac{1}{2}} = -5^{\frac{1}{2}}.$$

We can therefore transform the relation II into

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i}{k}(hn-h'')} = \sum_{h \bmod k}{}' \omega_{h,k} e^{-\frac{2\pi i h}{k} 5^*(n+1)}$$

or

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(5)} e^{-\frac{2\pi i}{k}(hn-h'')} = \sum_{h \bmod k}{}' \omega_{5h,k} e^{-\frac{2\pi i h}{k}(n+1)}. \tag{7.1}$$

Now, for odd k, we take the value of $\Omega_{h,k}^{(5)}$ from (6.23) and the value of $\omega_{5h,k}$ from (6.11) and find

$$\frac{\Omega_{h,k}^{(5)} e^{\frac{2\pi i h''}{k}}}{\omega_{5h,k} e^{-\frac{2\pi i h}{k}}} = \exp\left\{-\pi i\left(\frac{1}{12}\left(k - \frac{1}{k}\right)(-48h - 24h'' - 120h^2 h'') - \frac{2h'' + 2h}{k}\right)\right\}$$

$$= \exp\left\{-\frac{2\pi i}{k} h(1 + 5hh'')\right\} = 1,$$

where the last equality is a consequence of the definition $5hh'' \equiv -1 \pmod{k}$. Thus we have

$$\Omega_{h,k}^{(5)} e^{-\frac{2\pi i}{k}(hn-h'')} = \omega_{5h,k} e^{-\frac{2\pi i h}{k}(n+1)},$$

which proves (7.1) for odd k.

For even k we use (6.3) and (6.12) and find again

$$\frac{\Omega_{h,k}^{(5)} e^{\frac{2\pi i h''}{k}}}{\omega_{5h,k} e^{-\frac{2\pi i h}{k}}} = \exp\left\{-\pi i\left(\frac{1}{4}(24hk + 24h)\right.\right.$$

$$\left.\left. + \frac{1}{12}\left(k - \frac{1}{k}\right)(-48h - 24h'' - 120h^2 h'') - \frac{2h'' + 2h}{k}\right)\right\} = 1$$

which completes the proof of (7.1).

8. The relation III can be written

$$A_{\beta^{\beta}k_1}(5n+4) = 0$$

with $\beta > 1$, $5 \nmid k$ and is, as such, an immediate consequence of Lemma 2.

Thus we have proved the three relations for the modulus 5 and hence have proved Ramanujan's identity (1.11).

9. We now turn to the proof of the four relations concerning the modulus 7. The proof of I′ is analogous to that for I so we shall omit all detail and merely state the results. In this case we have $7 \nmid k$ and

$$7hh'' \equiv -1, \qquad 7h'' \equiv h' \pmod{k}$$

and we obtain

$$\Omega_{h,k}^{(7)} = \left(\frac{-7h}{k}\right) \exp\left\{-\pi i\left(\frac{1}{4}(k-1) + \frac{1}{12}\left(k - \frac{1}{k}\right)(-82h - 49h'' - 287h^2h'')\right)\right\}$$

for odd k, and

$$\Omega_{h,k}^{(7)} = \left(\frac{-k}{7h}\right) \exp\left\{-\pi i\left(\frac{1}{4}(2 + 41hk + 41h) + \frac{1}{12}\left(k - \frac{1}{h}\right)(-82h - 49h'' - 287h^2h'')\right)\right\}$$

for even k. In both cases we get

$$\Omega_{7h,k}^{(7)} = \omega_{h,k} e^{-\frac{24\pi i h}{k}}$$

from which we have

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(7)} e^{-\frac{2\pi i h(n-1)}{k}} = \sum_{h \bmod k}{}' \Omega_{7h,k}^{(7)} e^{-\frac{14\pi i h(n-1)}{k}} = \sum_{h \bmod k}{}' \omega_{h,k} e^{-\frac{2\pi i h}{k}(7n+5)}$$

which proves I′.

10. At this point it is convenient to prove the relation IV′. From the definition (2.52) we have, for $7 \nmid k$,

$$\Omega_{h,k}^{(3)} = \frac{\omega_{h,k}^{4}}{\omega_{7h,k}^{3}}, \qquad \Omega_{h,k}^{(7)} = \frac{\omega_{h,k}^{8}}{\omega_{7h,k}^{7}}$$

and hence

$$\frac{\Omega_{h,k}^{(3)}}{\Omega_{h,k}^{(7)} e^{\frac{2\pi i(h+h'')}{k}}} = \frac{\omega_{7h,k}^{4}}{\omega_{h,k}^{4}} e^{-\frac{2\pi i(h+h'')}{k}}.$$

For odd k we use (6.11) to obtain

$$\frac{\omega_{7h,k}^4}{\omega_{h,k}^4} e^{-\frac{2\pi i(h+h'')}{k}}$$

$$= \exp\left\{-\pi i\left(\frac{1}{3}\left(k-\frac{1}{k}\right)(12h+6h''+42h^2h'')+\frac{2(h+h'')}{k}\right)\right\} \tag{10.11}$$

$$= \exp\left\{\frac{2\pi ih}{k}(1+7hh'')\right\} = 1.$$

For even k we use (6.12) and find

$$\frac{\omega_{7h,k}^4}{\omega_{h,k}^4} e^{-\frac{2\pi i(h+h'')}{k}} = \exp\left\{-\pi i\left(-6hk-6h\right.\right.$$

$$\left.\left.+\frac{1}{3}\left(k-\frac{1}{k}\right)(12h+6h''+42h^2h'')+\frac{2(h+h'')}{k}\right)\right\} = 1 \tag{10.12}$$

as before. Hence, in both cases, we have

$$\Omega_{h,k}^{(3)} = \Omega_{h,k}^{(7)} e^{\frac{2\pi i(h+h'')}{k}} \tag{10.13}$$

which clearly proves IV′.

11. Equation (10.13) reduces II′ to

$$-\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = \frac{1}{\sqrt{7}} A_{7k}(7n+5) \tag{11.11}$$

where we have used the notation of (5.1). Now Lemma 1 yields the result

$$A_{7k}(7n+5) = A_7(5)A_k(7^*(n+1))$$

for $7 \nmid k$ and $7 \cdot 7^* \equiv 1 \pmod{k}$. Also, by Lehmer's Theorem 5, we find

$$A_7(5) = \left(\frac{3}{7}\right)\sqrt{7} = -\sqrt{7}.$$

Therefore (11.11) can be written as

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = A_k(7^*(n+1))$$

$$= \sum_{h \bmod k}{}' \omega_{h,k} e^{-\frac{2\pi ih}{k}7^*(n+1)}$$

or as

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = \sum_{h \bmod k}{}' \omega_{7h,k} e^{-\frac{2\pi ih}{k}(n+1)}. \tag{11.12}$$

Now, however, by (2.52), (10.11) and (10.12) we have

$$\Omega^{(3)}_{h,k} e^{-\frac{2\pi i}{k}(hn-h'')} = \frac{\omega^4_{h,k}}{\omega^3_{7h,k}} e^{-\frac{2\pi i}{k}(hn-h'')}$$

$$= \omega_{7h,k} e^{-\frac{2\pi i h}{k}(n+1)}$$

which proves (11.12) and hence II′.

12. The relation III′ is an immediate consequence of Lemma 2. This completes the proof of the four relations and hence proves Ramanujan's identity (1.12).

UNIVERSITY OF PENNSYLVANIA,
PHILADELPHIA, PA.

For odd k we use (6.11) to obtain

$$\frac{\omega_{7h,k}^4}{\omega_{h,k}^4} e^{-\frac{2\pi i(h+h'')}{k}}$$

$$(10.11) \qquad = \exp\left\{-\pi i\left(\frac{1}{3}\left(k-\frac{1}{k}\right)(12h+6h''+42h^2h'')+\frac{2(h+h'')}{k}\right)\right\}$$

$$= \exp\left\{\frac{2\pi ih}{k}(1+7hh'')\right\} = 1.$$

For even k we use (6.12) and find

$$(10.12) \qquad \frac{\omega_{7h,k}^4}{\omega_{h,k}^4} e^{-\frac{2\pi i(h+h'')}{k}} = \exp\left\{-\pi i\left(-6hk-6h+\frac{1}{3}\left(k-\frac{1}{k}\right)(12h+6h''+42h^2h'')+\frac{2(h+h'')}{k}\right)\right\} = 1$$

as before. Hence, in both cases, we have

$$(10.13) \qquad \Omega_{h,k}^{(3)} = \Omega_{h,k}^{(7)} e^{\frac{2\pi i(h+h'')}{k}}$$

which clearly proves IV′.

11. Equation (10.13) reduces II′ to

$$(11.11) \qquad -\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = \frac{1}{\sqrt{7}} A_{7k}(7n+5)$$

where we have used the notation of (5.1). Now Lemma 1 yields the result

$$A_{7k}(7n+5) = A_7(5)A_k(7^*(n+1))$$

for $7 \nmid k$ and $7 \cdot 7^* \equiv 1 \pmod{k}$. Also, by Lehmer's Theorem 5, we find

$$A_7(5) = \left(\frac{3}{7}\right)\sqrt{7} = -\sqrt{7}.$$

Therefore (11.11) can be written as

$$\sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = A_k(7^*(n+1))$$

$$= \sum_{h \bmod k}{}' \omega_{h,k} e^{-\frac{2\pi ih}{k}7^*(n+1)}$$

or as

$$(11.12) \qquad \sum_{h \bmod k}{}' \Omega_{h,k}^{(3)} e^{-\frac{2\pi i}{k}(hn-h'')} = \sum_{h \bmod k}{}' \omega_{7h,k} e^{-\frac{2\pi ih}{k}(n+1)}.$$

Now, however, by (2.52), (10.11) and (10.12) we have

$$\Omega^{(3)}_{h,k}\, e^{-\frac{2\pi i}{k}(hn-h'')} = \frac{\omega^{4}_{h,k}}{\omega^{3}_{7h,k}}\, e^{-\frac{2\pi i}{k}(hn-h'')}$$

$$= \omega_{7h,k}\, e^{-\frac{2\pi i h}{k}(n+1)}$$

which proves (11.12) and hence II′.

12. The relation III′ is an immediate consequence of Lemma 2. This completes the proof of the four relations and hence proves Ramanujan's identity (1.12).

UNIVERSITY OF PENNSYLVANIA,
PHILADELPHIA, PA.

ANNALS OF MATHEMATICS
Vol. 40, No. 2, April, 1939

MODULARITY IN THE THEORY OF LATTICES*

BY L. R. Wilcox

(Received April 8, 1938)

Introduction

The theory of lattices has been developed by Dedekind[1] and during recent years by J. von Neumann,[2] G. Birkhoff,[3] O. Ore,[4] K. Menger[5] and others. Most of the results, however, pertain to modular lattices. It is the purpose of the present work to generalize portions of the theory of modular lattices to arbitrary lattices when such generalization is possible (Section 1), or to lattices satisfying weak modular hypotheses (Sections 2–5).

The first section is devoted to a study of a binary relation called *modularity* (cf. Definition 1.1(a)) defined in a lattice L, and its connections with a theory of independence for L. Consequences of the assumption that the relation of *independence* (Definition 1.1(b)) is symmetric are contained in Section 2. In Section 3 the concept of semi-modular lattice is defined, and a general example including affine geometry is given. It is shown in Section 4 that a theory of such lattices may be obtained which closely approximates the Dedekind theory of modular lattices; in particular the Jordan theorem for principal chains is proved. Section 5 deals with semi-modular lattices satisfying both descending and ascending chain conditions. It is found that a dimension function exists having the essential properties of dimensionality for linear subspaces of a finite-dimensional (non-centered) affine space. Our results generalize those obtained through entirely different methods by K. Menger[6] for lattices satisfying our axioms and in addition an axiom of complementation.

By a *lattice* we shall mean a set L of elements $a, b, c, \cdots$, *partially ordered* by a binary relation $<$ defined for all pairs of elements of L (i.e., the relation $<$ satisfies (1) never $a < a$, (2) $a \leqq b$, $b \leqq a$ implies $a = b$, and (3) $a \leqq b$, $b \leqq c$ implies $a \leqq c$), with the further property that every two elements $a, b \in L$ possess a *least upper bound* $a + b$ and a *greatest lower bound* ab, characterized respectively by the properties $c \geqq a, b$ if and only if $c \geqq a + b$, and $c \leqq a, b$

* Presented to the American Mathematical Society Dec. 29, 1937.

[1] Ges. Werke, Braunschweig, 1932, Vol. II, pp. 236–271.

[2] Proc. Nat. Acad. Sci. 22 (1936), pp. 93–100; Proc. Nat. Acad. Sci. 23 (1937), pp. 17–22. Cf. also mimeographed and planographed lecture notes, "Continuous Geometry," Princeton, 1935–1937.

[3] Proc. Camb. Phil. Soc. 29 (1933), pp. 441–464; Proc. Camb. Phil. Soc. 31 (1935), pp. 433–454; Annals of Math. 36 (1935), pp. 743–748.

[4] Annals of Math. 36 (1935), pp. 406–437.

[5] Annals of Math. 37 (1936), pp. 456–482.

[6] Annals of Math. 37 (1936), pp. 464, 465.

if and only if $c \leqq ab$. A *zero* is an element $0 \in L$ with the property $0 \leqq a$ for every $a \in L$; a *unit* is an element $1 \in L$ with the property $1 \geqq a$ for every $a \in L$. A zero (unit) is unique if existent. Throughout the paper the existence of 0 is assumed, but except where explicit mention of 1 is made, it is not supposed that 1 exists. We assume the reader to be familiar with the elements of the theory of lattices.

We follow the standard set-theoretical notations as outlined, for example, in a paper of F. J. Murray and J. von Neumann,[7] with the exception that we denote for a function $f(x)$ the set of all elements $f(x)$ with the property $\epsilon(x)$ by $[f(x); \epsilon(x)]$, reserving the notation $(f(x); \epsilon(x))$ for the system of such elements; e.g., $(a_i; i = 1, 2, \cdots)$ denotes the sequence rather than the set of elements a_i. The empty set is denoted by Θ.

1. Modularity and the theory of independence

In this section we deal with a given lattice L with zero 0.

DEFINITION 1.1: Let $b, c \in L$.

(a) $(b, c)M$ *(read* (b, c) *modular) means* $(a + b)c = a + bc$ *for every* $a \leqq c$;

(b) $(b, c)\perp$ *(read* (b, c) *independent) means* $(b, c)M$, $bc = 0$;

(c) $b \parallel c$ *(read* b *parallel to* c*) means* $bc = 0$, $(b, c)\bar{M}$ *(*$\bar{M}$ *being the negation of the relation* M*).*

COROLLARY:

(a) *If* $b \leqq c$ *or* $b \geqq c$, *then* $(b, c)M$.

(b) *If* $b \in L$, *then* $(0, b)M$, $(b, 0)M$, $(b, b)M$, *and* $(1, b)M$, $(b, 1)M$ *when* 1 *exists.*

PROOF: Suppose $b \leqq c$ and let $a \leqq c$. Then $a + b \leqq c$, whence $(a + b)c = a + b = a + bc$. Suppose $b \geqq c$ and let $a \leqq c$. Then $a \leqq c \leqq b$, and $(a + b)c = c = bc = a + bc$. Thus $(b, c)M$ in either case, and (a) is established. Obviously (b) follows from (a).

REMARK: It should be observed that $(b, c)M$ is equivalent to the apparently weaker statement $(a + b)c \leqq a + bc$ for every $a \leqq c$, since it follows from $(a + b)c \geqq a$, $(a + b)c \geqq bc$ that $(a + b)c \geqq a + bc$. This fact will be used freely in what follows.

DEFINITION 1.2:

(a) $a > b$ *(read* a *prime over* b*) means* $a > b$ *together with* $a \geqq c \geqq b$ *implies* $a = c$ *or* $b = c$;

(b) $a < b$ *means* $b > a$;

(c) P *is the set of all* $a \in L$ *such that* $a > 0$. *The elements of* P *are called points and are denoted by* $p, q, r, \cdots$.

COROLLARY: *If* $p \in P$, $b \in L$, *then* $(b, p)M$.

PROOF: Let $a \leqq p$. Then $a = 0$ or $a = p$. If $a = 0$, then $(a + b)p = bp = a + bp$; if $a = p$, then $(a + b)p = (p + b)p = p = p + bp = a + bp$. Thus $(b, p)M$.

REMARK: It is in general false that $(p, c)M$ if $p \in P$, $c \in L$. For, consider the lattice consisting of $0, p, q, c, 1$ with $0 < q < c < 1$, $0 < p < 1$, $p \not< q$, $p \not< c$,

[7] Annals of Math. 37 (1936), p. 126.

$q \not< p$, $c \not< p$. Hence $q < c$, but $(q + p)c = c$, $q + pc = q \neq c$. This example shows also that the relation M is not necessarily symmetric.

LEMMA 1.1: *If* $(a, b)M$, $ab \leqq c$, *then* $(ac, b)M$.

PROOF: Let $d \leqq b$. Then $(d + ac)b \leqq (d + a)b = d + ab = d + acb$. Hence $(d + ac)b \leqq d + acb$, and $(ac, b)M$.

COROLLARY: *If* $(a, b)\perp$, $a' \leqq a$, *then* $(a', b)M$.

PROOF: Since $ab = 0$ we have $ab \leqq a'$, whence Lemma 1.1 yields $(aa', b)M$, i.e., $(a', b)M$.

THEOREM 1.1: *If* $(a, b)\perp$, $a' \leqq a$, $b' \leqq b$, *then* $(a', b')\perp$.

PROOF: Clearly $a'b' \leqq a'b \leqq ab = 0$, whence $a'b' = a'b = 0$. It remains to show $(a', b')M$. Now by Lemma 1.1 $(a', b)M$; hence for every $b'' \leqq b'$

$$(b'' + a')b' = (b'' + a')bb' = (b'' + a'b)b' = b''b' = b'' = b'' + a'b',$$

and the modularity of (a', b') is established.

LEMMA 1.2: *If* $(a, b)M$ *and if* $(c, a + b)M$, $c(a + b) \leqq a$, *then* $(c + a, b)M$ *and* $(c + a)b = ab$.

PROOF: Let $b' \leqq b$. Then

$$(b' + (c + a))b = (b' + (c + a))(a + b)b = ((b' + a) + c)(a + b)b$$
$$= (b' + a + c(a + b))b = (b' + a)b = b' + ab \leqq b' + (c + a)b,$$

whence $(b' + (c + a))b \leqq b' + (c + a)b$, and $(c + a, b)M$. Moreover,

$$(c + a)b = (a + c)(a + b)b = (a + c(a + b))b = ab,$$

and the proof is complete.

THEOREM 1.2: *If* $(a, b)M$ *and* $(c, a + b)\perp$, *then* $(c + a, b)M$ *and* $(c + a)b = ab$.

PROOF: This is trivial from Lemma 1.2 since $c(a + b) = 0 \leqq a$.

THEOREM 1.3: *If* $(a, b)M$ *and* $c \leqq b$, *then* $(c + a, b)M$.

PROOF: Let $b' \leqq b$. Then

$$(b' + (c + a))b = ((b' + c) + a)b = b' + c + ab = b' + (c + a)b,$$

whence $(c + a, b)M$.

COROLLARY: *If* $(a, b)\perp$ *and if* $c \leqq b$, *then* $(c + a, b)M$ *and* $(c + a)b = c$.

PROOF: This is obvious from Theorem 1.3.

REMARK: Theorems 1.2 and 1.3 express essential closure properties of the set of all modular pairs (a, b), stating that the relation $(a, b)M$ is not destroyed when a is replaced by $a + c$, where either $(c, a + b)\perp$ or $c \leqq b$.

LEMMA 1.3: *If* $(a, b)\perp$ *and* $(c, a + b)\perp$, *then* $(c + a, b)\perp$.

PROOF: By Theorem 1.2, $(c + a, b)M$; moreover $ab = 0$ yields $(c + a)b = 0$, whence $(c + a, b)\perp$.

DEFINITION 1.3: Let $n = 1, 2, \cdots$ *and* $a_1, \cdots, a_n$ *be given. We define* $(a_1, \cdots, a_n)\perp$ *to mean* $(\sum (a_i; i \in S), \sum (a_i; i \in T))\perp$ *for every* S, $T \perp [1, \cdots, n]$ *for which* $j \in S$, $k \in T$ *implies* $j < k$.

COROLLARY: *Let* $(a_1, \cdots, a_n)\perp$.

(a) *If* $a_i \neq 0$ *for* $i = 1, \cdots, n$, *then* $a_i \neq a_j$ *for* $i \neq j$;

(b) *if* $1 \leqq k_1 < k_2 < \cdots < k_t \leqq n$, *then* $(a_{k_1}, \cdots, a_{k_t})\perp$;
(c) *if* $a_i' \leqq a_i$ *for* $i = 1, \cdots, n$, *then* $(a_1', \cdots, a_n')\perp$.

PROOF: Part (b) is immediate, and (a) is trivial. To prove (c) we observe that if $S, T \subset [1, \cdots, n]$ and if $j \in S$, $k \in T$ implies $j < k$, then $(\sum (a_i'; i \in S), \sum (a_i'; i \in T))\perp$ follows from $(\sum (a_i; i \in S), \sum (a_i; i \in T))\perp$, $\sum (a_i'; i \in S) \leqq \sum (a_i; i \in S)$, $\sum (a_i'; i \in T) \leqq \sum (a_i; i \in T)$ by virtue of Theorem 1.1.

REMARK: The definition of independence of n elements when $n = 2$ obviously agrees with that given in Definition 1.1(b). When $n = 1$ Definition 1.3 is vacuously satisfied, and so every one-element system is independent. It should be observed that in proving $(a_1, \cdots, a_n)\perp$ one need never consider the case $S = \Theta$ or $T = \Theta$, since $\sum (a_i; i \in \Theta) = 0$, and $(0, d)M$, $(d, 0)M$ by Corollary Definition 1.1.

THEOREM 1.4: *Let* $n = 1, 2, \cdots$ *and* $a_1, \cdots, a_n$ *be given. Then* $(a_1, \cdots, a_n)\perp$ *if and only if* $(a_i, a_{i+1} + \cdots + a_n)\perp$ *for* $i = 1, \cdots, n-1$.

PROOF: The forward implication is evident. Suppose now that $(a_i, a_{i+1} + \cdots + a_n)\perp$ for $i = 1, \cdots, n-1$. We shall prove by induction that $(a_k, \cdots, a_n)\perp$ for $k = n, n-1, \cdots, 1$, whence the desired result follows by putting $k = 1$. For $k = n$ this is obvious, and for $k = n-1$ it follows from our hypothesis. Hence let $k = n-1, n-2, \cdots, 2$, and suppose that $(a_k, \cdots, a_n)\perp$. In order to show that $(a_{k-1}, a_k, \cdots, a_n)\perp$, let $S = [i_1, \cdots, i_r]$, $T = [j_1, \cdots, j_s]$, where $S, T \subset [k-1, k, \cdots, n]$, the i_ρ are distinct, the j_σ are distinct, and $i_\rho < j_\sigma$. For convenience let $i_1 < i_2 < \cdots < i_r < j_1 < \cdots < j_s$. If $i_1 \geqq k$, then $(\sum (a_j; j \in S), \sum (a_k; k \in T))\perp$ by the induction hypothesis. Suppose $i_1 < k$, i.e., $i_1 = k-1$. Define

$$a \equiv a_{i_1} = a_{k-1}, \quad b \equiv a_{i_2} + a_{i_3} + \cdots + a_{i_r}, \quad c \equiv a_{j_1} + a_{j_2} + \cdots + a_{j_s}.$$

Since $[i_2, \cdots, i_r, j_1, \cdots, j_s] \subset [k, \cdots, n]$, we have $(b, c)\perp$ by the induction hypothesis. Moreover, since $(a_{k-1}, a_k + \cdots + a_n)\perp$, and since $b + c \leqq a_k + \cdots + a_n$, it follows from Theorem 1.1 that $(a, b + c)\perp$. By Lemma 1.3 $(a + b, c)\perp$, i.e., $(\sum (a_j; j \in S), \sum (a_k; k \in T))\perp$. Thus Definition 1.3 is satisfied in either case, and $(a_{k-1}, a_k, \cdots, a_n)\perp$. This completes the proof.

LEMMA 1.4: *Let* $n = 1, 2, \cdots$. *If* $(a_1, \cdots, a_n)\perp$, *and* $S, T \subset [1, \cdots, n]$ *are disjoint, then*

$$\sum (a_i; i \in S) \sum (a_i; i \in T) = 0. \tag{1.1}$$

PROOF: Let $S = [i_1, \cdots, i_r]$, $T = [j_1, \cdots, j_s]$. If $S, T \subset [n-1, n]$, then (1.1) obviously holds. Suppose that $k = n-1, n-2, \cdots, 2$ and that (1.1) holds for $S, T \subset [k, k+1, \cdots, n]$; we shall show that (1.1) holds for $S, T \subset [k-1, k, \cdots, n]$, whence the desired result holds by induction. Without loss of generality it may be assumed that $i_1 < j_1$ and $i_1 < \cdots < i_r, j_1 < \cdots < j_s$. If $i_1 > k-1$, then the induction hypothesis yields our result. Hence let $i_1 = k-1$. Then

$$a_{j_1} + a_{j_2} + \cdots + a_{j_s} \leqq a_k + a_{k+1} + \cdots + a_n, \text{ and } (a_{k-1}, a_k + a_{k+1} + \cdots + a_n)\perp,$$

whence

$$\sum (a_i; i \in S) \sum (a_i; i \in T) = (a_{i_1} + \cdots + a_{i_r})(a_k + \cdots + a_n)(a_{j_1} + \cdots + a_{j_s})$$
$$= (a_{i_2} + \cdots + a_{i_r})(a_{j_1} + \cdots + a_{j_s}) = 0,$$

the last equality holding by the induction hypothesis, since $[i_2, \cdots, i_r]$, $[j_1, \cdots, j_s] \subset [k, \cdots, n]$. This completes the proof.

LEMMA 1.5: *Let* $n = 1, 2, \cdots$ *and* $(a_1, \cdots, a_n)\perp$. *Moreover, let* $S, T \subset [1, \cdots, n]$ *be such that* $i < j$ *for* $i \in (S - ST), j \in T$. *Then*

$$\sum (a_i; i \in S) \sum (a_i; i \in T) = \sum (a_i; i \in ST).$$

PROOF: Since $(a_1, \cdots, a_n)\perp$, we have clearly $(\sum (a_i;\ i \in (S - ST)), \sum (a_i; i \in T))\perp$. Hence

$$\sum (a_i; i \in S) \sum (a_i; i \in T)$$
$$= (\sum (a_i; i \in ST) + \sum (a_i; i \in (S - ST))) \sum (a_i; i \in T) = \sum (a_i; i \in ST).$$

THEOREM 1.5: *Let* $n = 1, 2, \cdots$ *and* $(a_1, \cdots, a_n)\perp$. *Moreover, let* $S, T \subset [1, \cdots, n]$ *satisfy any one of the conditions* (a) $S \subset T$, (b) $S \supset T$, (c) $ST = \theta$, (d) $i < j$ *for* $i \in (S - ST), j \in T$, (e) $i < j$ *for* $i \in (T - ST), j \in S$. *Then*

$$\sum (a_i; i \in S) \sum (a_i; i \in T) = \sum (a_i; i \in ST). \tag{1.2}$$

PROOF: The result is trivial in case either (a) or (b) holds. When (c) holds, (1.2) from (1.1) in Lemma 1.4. If (d) holds, the result is that of Lemma 1.5, and if (e) holds, we apply the preceding case with S and T interchanged.

REMARK: If the modular axiom is assumed (i.e., if always $(a, b)M$), then (1.2) is obtained without restrictions on S, T (cf. Theorem 1.7 together with Theorem 2.1 for a generalization of this fact). However, Theorem 1.5 gives the strongest result that can be obtained without restrictions on the relation M, as is shown by examples.

DEFINITION 1.4: *Let* $n = 1, 2, \cdots$ *and* $a_1, \cdots, a_n$ *be given. We define* $(a_1, \cdots, a_n)\perp_s$ (*read* (a_i) *symmetrically independent*) *to mean* $(a_{j_1}, \cdots, a_{j_n})\perp$ *for every permutation* $i \to j_i$ *of the set of integers* $[1, \cdots, n]$.

REMARK: It is clear that the relation $\perp_s$ is symmetric; moreover, $(a_1, \cdots, a_n)\perp_s$ clearly implies $(a_1, \cdots, a_n)\perp$, while the converse is not generally true, as is shown by examples.

THEOREM 1.6: *If* $(a_1, \cdots, a_n)\perp_s$, *then* $(a_i, \sum_{j \neq i} a_j)\perp$ *for every* $i = 1, \cdots, n$, *and conversely.*

PROOF: If $(a_1, \cdots, a_n)\perp_s$, then $(a_i, a_2, \cdots, a_{i-1}, a_1, a_{i+1}, \cdots, a_n)\perp$, whence the desired condition holds. Conversely, let $i \to j_i$ be any permutation of $[1, \cdots, n]$ and define $b_i \equiv a_{j_i}$. Our hypothesis then yields that $(b_i, \sum_{k \neq i} b_k)\perp$ for $i = 1, \cdots, n$. Hence $(b_i, \sum_{k=i+1}^{n} b_k)\perp$ for $i = 1, \cdots, n-1$, and it follows from Theorem 1.4 that $(b_1, \cdots, b_n)\perp$. Thus $(a_{j_1}, \cdots, a_{j_n})\perp$, and the proof is complete.

REMARK: The condition $(a_i, \sum_{j \neq i} a_j)\perp$ $(i = 1, \cdots, n)$ appears in the theory of independence for modular lattices as a necessary and sufficient condition for $(a_1, \cdots, a_n)\perp$; Theorem 1.6 is thus a generalization of this result.

THEOREM 1.7: *Let* $(a_1, \cdots, a_n)\perp_s$, *and let* $S, T \subset [1, \cdots, n]$. *Then*

$$\sum (a_i; i \in S) \sum (a_i; i \in T) = \sum (a_i; i \in ST).$$

PROOF: Let $S - ST = [i_1, \cdots, i_r]$, $T - ST = [j_1, \cdots, j_s]$, $ST = [k_1, \cdots, k_t]$, where the i_ρ are mutually distinct, as are similarly the j_σ and the k_τ. Now there exists a permutation of $[1, \cdots, n]$ which brings about $i_1 < \cdots < i_r < j_1 < \cdots < j_s < k_1 < \cdots < k_t$, without destroying the relation $(a_1, \cdots, a_n)\perp_s$. We may thus suppose that this permutation has been effected and apply Lemma 1.5, obtaining the desired result.

We turn now to the consequences of a mild assumption concerning the relation $(a, b)M$.

2. Theory of symmetric lattices

Throughout this section we shall assume Axiom A (stated below) and develop the stronger theory of independence resulting from this assumption. The term *symmetric* is applied to lattices satisfying Axiom A, since for them the relation $(a, b)\perp$ is symmetric.

AXIOM A: *If* $(a, b)\perp$, *then* $(b, a)M$.

COROLLARY: *The relation* $(a, b)\perp$ *is symmetric, or equivalently,* $(a, b)\perp$ *if and only if* $(a, b)\perp_s$.

LEMMA 2.1: *If* $(b, a, c)\perp$, *then* $(a, b, c)\perp$.

PROOF: By the hypothesis, $(b, a)\perp$ and $(b + a, c)\perp$. Hence by Axiom A $(b, a)\perp$, $(c, b + a)\perp$. Now Lemma 1.3 applies, yielding $(c + b, a)\perp$, whence also $(a, b + c)\perp$ by Axiom A. Clearly $(b, c)\perp$; therefore $(a, b, c)\perp$ by Theorem 1.4.

THEOREM 2.1: *If* $n = 1, 2, \cdots$, *and* $(a_1, \cdots, a_n)\perp$, *then* $(a_1, \cdots, a_n)\perp_s$, *and conversely.*

PROOF: The converse is trivial. To prove the forward implication let $i = 1, \cdots, n$ be given and define

$$a \equiv a_i, \qquad b \equiv \sum_{j=1}^{i-1} a_j, \qquad c \equiv \sum_{j=i+1}^{n} a_j.$$

Clearly $(a, c)\perp$, $(b, a + c)\perp$, whence $(b, a, c)\perp$ by Theorem 1.4. Thus Lemma 2.1 yields $(a, b, c)\perp$, i.e., $(a_i, \sum_{j \neq i} a_j)\perp$. This holds for $i = 1, \cdots, n$, whence $(a_1, \cdots, a_n)\perp_s$ by Theorem 1.6.

COROLLARY: *If* $a_1, \cdots, a_n$ *are given, then* $(a_1, \cdots, a_n)\perp$ *if and only if* $(\sum (a_i; i \in S), \sum (a_i; i \in T))\perp$ *for every disjoint pair* $S, T \subset [1, \cdots, n]$.

PROOF: Suppose $(a_1, \cdots, a_n)\perp$, i.e., $(a_1, \cdots, a_n)\perp_s$, and let $S = [i_1, \cdots, i_s]$, $T = [j_1, \cdots, j_t]$, where the i_σ are mutually distinct, as are also the j_τ. Clearly there exists a permutation of $[1, \cdots, n]$ that effects $i_1 < i_2$

$< \cdots < i_s < j_1 < \cdots < j_t$ without destroying $(a_1, \cdots, a_n)\perp_s$. Hence we may suppose this permutation to have been performed and infer the desired conclusion by Definition 1.3. The converse is trivial.

REMARK: The condition in the corollary just proved appears in the theory of modular lattices as the definition of $(a_1, \cdots, a_n)\perp$. It should be noted that Theorem 2.1 yields the symmetry of the relation $(a_1, \cdots, a_n)\perp$.

LEMMA 2.2: *If* $(a, c)\perp$ *and* $(a + c, b + d)\perp$, *then* $(a + b)(c + d) = bd$.

PROOF: Since $(a, c)\perp$ and $(b + d, a + c)\perp$, we have by Theorem 1.4 that $(b + d + a, c)\perp$, whence $(c, a + b + d)\perp$. Hence

$$(a + b)(c + d) = (a + b)(a + b + d)(c + d) = (a + b)d$$
$$= (a + b)(b + d)d = (b + a(b + d))d = bd,$$

since $(a, b + d)\perp$.

THEOREM 2.2: *If* $(a_1, \cdots, a_n)\perp$, $(b_1, \cdots, b_m)\perp$ *and* $(a_1 + \cdots + a_n, b_1 + \cdots + b_m)\perp$, *then* $(a_1, \cdots, a_n, b_1, \cdots, b_m)\perp$.

PROOF: By Theorem 1.4 it suffices to prove

$$(a_i, a_{i+1} + \cdots + a_n + b_1 + \cdots + b_m)\perp \qquad (i = 1, \cdots, n) \tag{2.1}$$

and

$$(b_j, b_{j+1} + \cdots + b_m)\perp \qquad (j = 1, \cdots, m - 1). \tag{2.2}$$

Since (2.2) follows from $(b_1, \cdots, b_m)\perp$, we prove only (2.1). Define

$$a \equiv a_i, \qquad b \equiv a_{i+1} + \cdots + a_n, \qquad c \equiv b_1 + \cdots + b_m.$$

Clearly $(a, b)\perp$, $(a + b, c)\perp$, whence it follows that $(a, b, c)\perp$. Thus $(a, b + c)\perp$, i.e., (2.1) holds.

COROLLARY: *Suppose that* m *systems* $(a_{ji}; j = 1, \cdots, n_i)$, $i = 1, \cdots, m$ *are given. If* $(a_{ji}; j = 1, \cdots, n_i)\perp$ *for* $i = 1, \cdots, m$, *and if* $(b_1, \cdots, b_m)\perp$, *where* $b_i \equiv \sum_{j=1}^{n_i} a_{ji}$, *then* $(a_{ji}; j = 1, \cdots, n_i, i = 1, \cdots, m)\perp$.

PROOF: This follows immediately from Theorem 2.2 by induction on m.

REMARK: The result of Theorem 2.2 is an important property of independence in the theory of modular lattices; it is found to depend rather essentially on Axiom A, being false for a general lattice. It might seem that the result should hold in general when independence is replaced by symmetric independence, but this also is contradicted by examples.

3. Semi-modular lattices

In this and the succeeding section a theory of lattices satisfying Axiom A and Axiom B (stated below) is developed. This section is devoted to definitions, several fundamental properties and a general example of semi-modular lattices. The term *semi-modular* is applied to any lattice satisfying Axioms A and B. We assume henceforth that L is semi-modular.

AXIOM B: *If* $ab \neq 0$, *when* $(a, b)M$.

COROLLARY:
(a) $(a, b)M$ *if and only if* $(b, a)M$;
(b) *the statements* $(a, b)\bar{M}$, $a \| b$, $b \| a$ *are equivalent* (cf. Definition 1.1(c)).

PROOF: Immediate.

THEOREM 3.1: *If* $(a, c)M$, $(a + c, b + d)\perp$, *and* $(b, d)M$, *then* $(a + b)(c + d) = ac + bd$.

PROOF: Since $(a, c)M$, $(b + d, a + c)\perp$, we have $(a + b + d, c)M$ and $(a + b + d)c = ac$ by Theorem 1.2. By Corollary Axiom B $(c, a + b + d)M$. Now

$$(a + b)(c + d) = (a + b)(a + b + d)(c + d)$$
$$= (a + b)(d + (a + b + d)c) = (a + b)(d + ac).$$

But $(b, d)M$, $(a, b + d)\perp$, whence $(a + b, d)M$, i.e., $(d, a + b)M$, and $(a + b)d = bd$ by Theorem 1.2. Thus

$$(a + b)(c + d) = ac + d(a + b) = ac + bd,$$

and the proof is complete.

REMARK: The theorem just proved is a generalization of an important property of modular lattices. If $a = bd = 0$, the result holds for a general lattice as Theorem 1.4 and Lemma 1.4 show. If merely $ac = 0$, Axiom A is sufficient to prove it (cf. Lemma 2.2). However, the general form seems to depend essentially on the symmetry of the relation $(a, b)M$.

DEFINITION 3.1: We shall say that L *satisfies the ascending (descending) chain condition* in case there exists in L no denumerably infinite sequence $(a_i\,;\, i = 1, 2, \cdots)$ with $a_i < a_j (a_i > a_j)$ for $i < j$.

We interrupt the general theory with the description of an example of a semi-modular lattice.

Let Λ be a given lattice with 0 and 1, partially ordered by a relation $a < b$, and having operations $a \oplus b$, $a \otimes b$. Let Λ be *modular*, i.e., $(a, b)M$ for every $a, b \in \Lambda$, and *complemented*, i.e.,

(3.1) $a \leqq b$ implies the existence of $c \in \Lambda$ with $a \oplus c = b$, $a \otimes c = 0$.

Let $S \subset \Lambda$ be a fixed *ideal* with 0 deleted, i.e., a set with the properties

(3.2) $$a, b \in S \text{ implies } a \oplus b \in S,$$

(3.3) $$a \in S, \quad 0 < b \leqq a \text{ implies } b \in S,$$

(3.4) $$0 \text{ not in } S.$$

Define $L \equiv \Lambda - S$. Since $L \subset \Lambda$, L is partially ordered by the relation $a < b$. We shall prove that L is a lattice by introducing operations into L in accordance with the following definitions. Let $a, b \in L$; then define

$$a + b \equiv a \oplus b,$$

(3.5) $$ab \equiv \begin{cases} a \otimes b & \text{if } a \otimes b \in L \\ 0 & \text{if } a \otimes b \in S. \end{cases}$$

To show that $a + b$ is effective as least upper bound in L of a, b, it suffices to prove $a + b \in L$ if $a, b \in L$. If $a = 0$, then $a + b = 0 \oplus b = b$, whence $a + b \in L$. Suppose $a \neq 0$ and that $a + b \in S$. Then since $0 < a \leqq a + b$ we have by (3.3) that $a \in S$, contrary to $a \in L$. We shall now show directly that ab is effective as greatest lower bound of a, b. Clearly $ab \in L$ and $ab \leqq a, b$. Suppose now that $c \in L$, $c \leqq a, b$; then $c \leqq a \otimes b$. If $a \otimes b \in L$, then $c \leqq ab$, and the proof is complete in this case; if $a \otimes b \in S$, then $c = 0$, since otherwise $c \in S$ by (3.3), contrary to $c \in L$.

Before proving that L satisfies Axioms A and B, we establish an important lemma, viz.,

(3.6) $\qquad a \,||\, b$ *if and only if* $a \otimes b \in S$.

Let first $a \,||\, b$. If $a \otimes b \in L$, then $a \otimes b = ab = 0$. Now suppose $b' \in L$, $b' \leqq b$. If $b' = 0$, then $(b' + a)b = ab = b' + ab$; if $b' \neq 0$, then $(b' + a)b \geqq b' > 0$, and

$$(b' + a)b = (b' \oplus a) \otimes b = b' \oplus (a \otimes b) = b' + ab.$$

Thus in any case $(b' + a)b = b' + ab$, whence $(a, b)M$, contrary to $a \,||\, b$. Conversely, let $a \otimes b \in S$. Then $ab = 0$ by (3.5). Suppose a not parallel to b; hence $(a, b)M$. Now by (3.1) there exists $b' \in \Lambda$ such that

(3.7) $\qquad (a \otimes b) \oplus b' = b, \qquad (a \otimes b) \otimes b' = 0.$

If $b' \in S$, then $b \in S$ by (3.2), contrary to $b \in L$; therefore $b' \in L$. Now

$$b' + a = b' \oplus a = b' \oplus (a \otimes b) \oplus a = b \oplus a = b + a.$$

Hence $(b' + a)b = (b + a)b = b$, and $b' + ab = b'$. Since $b' \leqq b$ and since $(a, b)M$, we obtain $b' = b$, whence by (3.7) $a \otimes b = 0$, contrary to (3.4). This establishes $a \,||\, b$ and completes the proof of (3.6). An immediate corollary is

(3.8) $\qquad (a, b)\perp$ *if and only if* $a \otimes b = 0$.

For, let $(a, b)\perp$; then a is not parallel to b, and by (3.6) $a \otimes b \in L$, whence $a \otimes b = ab = 0$. Conversely, $a \otimes b = 0$ yields $a \otimes b \in L$ and $ab = a \otimes b = 0$; by (3.6) $a \otimes b \in L$ means a is not parallel to b, whence $(a, b)\perp$.

To prove Axiom A we need only to observe that by (3.8) $(a, b)\perp$ implies $a \otimes b = 0$, which in turn implies $b \otimes a = 0$; i.e., $(b, a)\perp$. Finally, to prove Axiom B, let $a, b \in L$, $ab \neq 0$, and let $b' \in L$, $b' \leqq b$; then $(b' + a)b \geqq ab > 0$, and $ab = a \otimes b$, whence

$$(b' + a)b = (b' \oplus a) \otimes b = b' \oplus (a \otimes b) = b' \oplus ab = b' + ab,$$

and $(a, b)M$. Thus it has been shown that L is a semi-modular lattice. If Λ is a *projective geometry*,[8] i.e., the lattice of all linear subsets of a finite-dimensional projective space Q, and if $h \in \Lambda$ is any *hyperplane*, i.e., satisfies $h \prec 1$, then $S \equiv [x; x \leqq h, x \neq 0]$ satisfies (3.2), (3.3), (3.4), and the lattice $L = \Lambda - S$

[8] Cf. G. Birkhoff, Annals of Math. 36 (1935), pp. 743–745.

is isomorphic to the lattice of all linear subsets of the (non-centered) affine space arising from Q by the deletion of the points of the hyperplane h. It is important to observe that L is non-modular (provided the dimension of Q is more than 1); indeed in this case the relation $a \parallel b$ is precisely the relation of parallelism as generally understood for affine spaces, since (3.6) shows that $a \parallel b$ means that a and b "intersect" in the deleted hyperplane h.

A preliminary study of examples of the type just described has been made by the author;[9] a more comprehensive discussion is to be given in the near future. It should be observed that the assumption (3.1) of complementation is stronger than necessary; it was needed only in the proof of the converse implication in (3.6), where the hypothesis

$$a \in L,\ b \in S,\ b \leqq a \text{ implies the existence of } c \in \Lambda \text{ with } b \oplus c = a,\ b \otimes c = 0$$

would be sufficient. Moreover, if in the condition above we demand that $c \in L$, then condition (3.2) may be deleted. However, these observations are not important for our present purposes.

4. The Jordan theorem for semi-modular lattices

Let us return to our semi-modular lattice L. We shall prove a sequence of lemmas culminating in the so-called Jordan theorem concerning principal chains (cf. Theorem 4.1). The lemmas are generalizations or modifications of those given by Dedekind[10] in the corresponding theory for modular lattices.

LEMMA 4.1: *If* $ab < a$, *then* $a + b > b$.

PROOF: Clearly $a + b \geqq b$; if $a + b = b$, then $ab = a$, contrary to $ab < a$. Hence $a + b > b$. Suppose now that $a + b \geqq e \geqq b$. Define $f \equiv ea$. Then $ab \leqq f \leqq a$, whence $f = ab$ or $f = a$ by the hypothesis. If $f = a$, then $ea = a$, i.e., $a \leqq e$, whence $a + b \leqq e + b = e$, and $a + b = e$ (since we had $a + b \geqq e$). On the other hand, let $f = ab$, i.e., $ea = ab$. Now if $ab \neq 0$, then $ea \neq 0$, and $(a, e)M$ by Axiom B. If $ea = ab = 0$, then $a > 0$, i.e., $a \in P$, and $(e, a)M$ by Corollary Definition 1.2. Thus also $(a, e)M$ by Axiom A, and we have

$$e = (b + a)e = b + ae = b + ab = b.$$

Therefore either $a + b = e$ or $b = e$, and we have established $a + b > b$ by Definition 1.2(a).

LEMMA 4.2: *If* $a + b > b$ *and if* $(b, a)M$, *then* $ab < a$.

PROOF: Clearly $ab \leqq a$; if $ab = a$, then $a + b = b$ contrary to $a + b > b$. Hence $ab < a$. Suppose now that $ab \leqq e \leqq a$. Define $f \equiv e + b$. Then $b \leqq f \leqq a + b$, whence $f = b$ or $f = a + b$ by the hypothesis. If $f = b$, then $e + b = b$, i.e., $e \leqq b$, whence $ab \geqq ae = e$, and $ab = e$. If on the other hand $f = a + b$, i.e., $e + b = a + b$, then

$$a = (a + b)a = (e + b)a = e + ab = e.$$

Hence either $ab = e$ or $a = e$, and $ab > a$.

[9] Bull. Amer. Math. Soc. 43 (1937), p. 183, Abstract 159.

[10] Werke, Vol. II, pp. 259–265.

REMARK: Lemma 4.2 is true independently of Axioms A and B, as is seen from the proof, while Lemma 4.1 is false for a general lattice.

LEMMA 4.3: *If $c \prec d$ and $e \in L$, then either $c + e = d + e$ or $c + e \prec d + e$.*

PROOF: Clearly $c + e = d + e$ or $c + e < d + e$. Suppose in the latter case that $c + e \leqq f \leqq d + e$. Then $c \leqq (c + e)d \leqq fd \leqq d$, whence $c \leqq df \leqq d$. If $df = 0$, then $c = 0$ and $d \in P$, and we may conclude $(d, f)M$. If $df \neq 0$, then $(d, f)M$ by Axiom B. Therefore $(d, f)M$ in either case, whence

$$e + df = (e + d)f = f,$$

since $e \leqq c + e \leqq f$. Now since $c \prec d$, we have $df = c$ or $df = d$. If $df = c$, then $e + c = e + df = f$; if $df = d$, then $e + d = e + df = f$. This shows that either $c + e = f$ or $d + e = f$, whence $c + e \prec d + e$.

LEMMA 4.4: *If $b \prec c$, $(a, c)M$ and $(a, b)M$, then either $a + b \prec a + c$ and $ab = ac$, or $a + b = a + c$ and $ab \prec ac$.*

PROOF: Define $d \equiv (b + a)c = b + ac$. Then $b \leqq d \leqq c$. Hence $d = b$ or $d = c$. Let us first suppose $d = b$. Then $b + ac = b$, i.e., $ac \leqq b$. Thus $ac \leqq ab \leqq ac$, and we obtain $ac = ab$. Moreover $d \prec c$, i.e., $(b + a)c \prec c$, whence by Lemma 4.1 $a + b + c \succ a + b$, or equivalently $a + c \succ a + b$. Thus we have established the first of the two alternatives in the conclusion of the lemma. Let us now suppose $d = c$. Then $(b + a)c = c$, i.e., $b + a \geqq c$. Thus $b + a \geqq a + c \geqq a + b$, whence $a + b = a + c$. Now $b \prec c = d = b + ac$. Since $(a, b)M$, $ab \leqq c$, we have $(ac, b)M$ by Lemma 1.1. Thus by Lemma 4.2 $acb \prec ac$, or equivalently $ab \prec ac$. This establishes the second alternative and completes the proof.

DEFINITION 4.1: *A principal increasing (decreasing) chain is a finite sequence $(a_1, \cdots, a_n)$ such that $a_i \prec a_{i+1} (a_i \succ a_{i+1})$ for $i = 1, \cdots, n - 1$; in either case the integer n is referred to as the length of the chain, and the elements $a_1, \cdots, a_n$ are called members of the chain.*

LEMMA 4.5: *If $(a_1, \cdots, a_n)$ is a principal increasing chain, and if $a_1 \leqq a \leqq a_n$, $(a, a_i)M$ for $i = 1, \cdots, n$, then there exists a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ having the same length as the given chain and containing a as a member, such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$.*

PROOF: If we form the products $a_i a$ and the sums $a_i + a$, then the following relations hold:

$$a_1 = a_1 a \leqq a_2 a \leqq \cdots \leqq a_n a = a = a_1 + a \leqq a_2 + a \leqq \cdots \leqq a_n + a = a_n.$$

Since by Lemma 4.4 either $a_i a = a_{i+1} a$ and $a_i + a \prec a_{i+1} + a$, or $a_i a \prec a_{i+1} a$ and $a_i + a = a_{i+1} + a$, it follows that for each $i = 1, \cdots, n - 1$ one and only one of the pairs $(a_i a, a_{i+1} a)$, $(a_i + a, a_{i+1} + a)$ satisfies the relation $=$ and the other the relation $\prec$. Hence if we delete from the system $(a_1 a, a_2 a, \cdots, a_n a = a_1 + a, a_2 + a, \cdots, a_n + a)$ each term $a_i a$ or $a_i + a$ equal to its predecessor, we shall have deleted exactly $n - 1$ terms. Since there are altogether $2n - 1$ terms, it follows that there will remain a system of exactly n terms. This system $(\bar{a}_1, \cdots, \bar{a}_n)$ is clearly a principal increasing chain of

length n; moreover $a = a_1 + a$ is one of its members. Finally, $\bar{a}_1 = a_1 a = a_1$, $\bar{a}_n = a_n + a = a_n$, and the proof is complete.

COROLLARY: *If $(a_1, \cdots, a_n)$ is a principal increasing chain with $a_1 \neq 0$, and if $a_1 \leqq a \leqq a_n$, then there exists a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ having length n and containing a as a member, such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$.*

PROOF: Since $a_1 \leqq a$, we have $a_1 a \neq 0$, whence $a_i a \geqq a_1 a > 0$, i.e., $a a_i \neq 0$ for $i = 1, \cdots, n$. Thus $(a, a_i)M$, and the desired conclusion follows from Lemma 4.5.

LEMMA 4.6: *Let L satisfy the descending chain condition* (cf. Definition 3.1). *If $(a_1, \cdots, a_n)$ is a principal increasing chain, and if $a_1 \leqq a \leqq a_n$, then there exists a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ having length n and containing a as a member, such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$.*

PROOF: If $a = a_1$, the result is trivial; hence let $a > a_1$. If $a_1 \neq 0$, the desired result holds by Corollary Lemma 4.5. Hence suppose $a_1 = 0$. Now from the descending chain condition we may infer that there exists $p \in P$ with $p \leqq a$. For otherwise it is possible to construct a sequence $(b_i; i = 1, 2, \cdots)$ with $b_1 > b_2 > \cdots$ by defining b_1 as a, b_2 as any element such that $0 < b_2 < b_1$, existent since b_1 is not in P, b_3 as any element such that $0 < b_3 < b_2$, existent since b_2 is not in P, etc.; the existence of the sequence (b_i) contradicts the descending chain condition. Now $(p, a_i)M$, whence by Lemma 4.5 there exists a principal increasing chain $(a_1', \cdots, a_n')$ with p as a member and such that $a_1' = 0$, $a_n' = a_n$. Since p is a member of $(a_1', \cdots, a_n')$, there exists $j = 1, \cdots, n$ such that $p = a_j'$. Obviously $j \neq 1$, whence $j \geqq 2$; but if $j > 2$, then $0 < a_2' < p$, contrary to $p > 0$. Hence $j = 2$, and our chain takes the form $(0, p, a_3', \cdots, a_n')$. For $i \geqq 3$ $a a_i' \geqq p > 0$, whence $(a, a_i)M$ by Axiom B. But $(a, 0)M$, $(a, p)M$, and therefore Lemma 4.5 applies and yields a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ containing a as a member and having the properties $\bar{a}_1 = a_1' = a_1$, $\bar{a}_n = a_n' = a_n$.

LEMMA 4.7: *Let L satisfy the descending chain condition. If $(a_1, \cdots, a_n)$ is a principal increasing chain, and if $a_1', \cdots, a_m'$ are elements satisfying*

$$a_1 \leqq a_1' < a_2' < \cdots < a_m' \leqq a_n,$$

then there exists a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ in which all the a_i' appear as members and such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$.

PROOF: Since $a_1 \leqq a_1' \leqq a_n$, application of Lemma 4.6 yields a principal increasing chain $(\bar{a}_1^{(1)}, \cdots, \bar{a}_n^{(1)})$ in which a_1' appears, let us say as $\bar{a}_j^{(1)}$, and such that $\bar{a}_1^{(1)} = a_1$, $\bar{a}_n^{(1)} = a_n$. Applying Lemma 4.6 to the chain $(\bar{a}_j^{(1)}, \cdots, \bar{a}_n^{(1)})$, we obtain, since $\bar{a}_j^{(1)} < a_2' \leqq \bar{a}_n^{(1)}$, a principal increasing chain $(\bar{a}_j^{(2)}, \cdots, \bar{a}_n^{(2)})$ in which a_2' appears and such that $\bar{a}_j^{(2)} = \bar{a}_j^{(1)} = a_1'$, $\bar{a}_n^{(2)} = \bar{a}_n^{(1)} = a_n$. If we define $\bar{a}_1^{(2)} \equiv \bar{a}_1^{(1)}, \cdots, \bar{a}_{j-1}^{(2)} \equiv \bar{a}_{j-1}^{(1)}$, then the principal chain $(\bar{a}_1^{(2)}, \cdots, \bar{a}_n^{(2)})$ contains a_1', a_2' as members and has the properties $\bar{a}_1^{(2)} = a_1$, $\bar{a}_n^{(2)} = a_n$. Proceeding thus inductively, we obtain finally a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ in which all the a_i' appear and such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$.

THEOREM 4.1: *Let L satisfy the descending chain condition. If m and n are*

given integers such that there exist principal increasing chains $(a_1, \cdots, a_n)$, $(a_1', \cdots, a_m')$ *with* $a_1 = a_1'$, $a_n = a_m'$, *then* $m = n$.

PROOF: Since $a_1 \leqq a_1' < \cdots < a_m' \leqq a_n$, we may apply Lemma 4.7 and obtain a principal increasing chain $(\bar{a}_1, \cdots, \bar{a}_n)$ in which all the a_i' appear and such that $\bar{a}_1 = a_1$, $\bar{a}_n = a_n$. Hence the set of the m elements a_i' is contained in the set of the n elements $\bar{a}_i$, whence it follows that $m \leqq n$. By symmetry $n \leqq m$. Thus $m = n$.

REMARK: Theorem 4.1 is our generalization of the Jordan theorem for principal chains. No attempt has been made to prove the theorem without the hypothesis that L satisfy the descending chain condition since no use will be made of the theorem except when this condition holds. However, it is clear that weaker conditions would suffice. For example, if L has the property that for each $a \in L$ there is an element $p \in P$ with $p \leqq a$, the proof of Lemma 4.6 would be valid as it stands, and hence Theorem 4.1 would also hold. This question is of little importance to us at present.

5. Semi-modular lattices satisfying the chain conditions

It is assumed in this section that L satisfies Axioms A and B, and in addition

AXIOM C: *L satisfies both ascending and descending chain conditions.*

We shall see that Theorem 4.1 implies the existence of an integer-valued dimension function having properties somewhat weaker than dimensionality in projective spaces but quite similar to the properties of dimensionality in affine spaces. It follows immediately from Axiom C that a unit $1 \in L$ must exist. The results of Lemmas 5.1 and 5.2 are well known,[11] but we include the proofs for the sake of completeness; these lemmas are consequences of Axiom C alone.

LEMMA 5.1: *Let $S \subset L$ be linearly ordered by the relation $<$, i.e., let S satisfy the condition*

$$a, b \in S \quad \textit{implies} \quad a \leqq b \textit{ or } a \geqq b. \tag{5.1}$$

Then S is finite.

PROOF: Suppose that S is infinite, and let T be any non-empty subset of S. It is clear that T has a smallest element $a \leqq b$ for every $b \in T$, since otherwise an infinite decreasing sequence may be constructed in T, contrary to the descending chain condition. Thus every non-empty set $T \subset S$ has a smallest element, whence S is well-ordered by the relation $<$. But S is infinite, and therefore there exists an increasing denumerable sequence $(a_i\,;\ i = 1, 2, \cdots)$ of elements in S, contrary to the ascending chain condition. This shows that S is finite.

LEMMA 5.2: *If $a \leqq b$, then there exists an integer $n = 1, 2, \cdots$ and a principal increasing chain $(a_1, \cdots, a_n)$ of length n such that $a_1 = a$, $a_n = b$.*

PROOF: The case $a = b$ is trivial; suppose $a < b$. If the conclusion of the

[11] Cf. Ore, Annals of Math. 36 (1935), p. 411.

lemma is false, then for every integer $n = 1, 2, \cdots$ and every sequence $(a_1, \cdots, a_n)$ with $a_1 < a_2 < \cdots < a_n$ it is true that the sequence $(a_1, \cdots, a_n)$ is not a principal chain. In particular, (a, b) is not a principal chain, whence there exists c with $a < c < b$. Moreover, (a, c, b) is not a principal chain, and there exists d with either $a < d < c$ or $c < d < b$. Proceeding inductively we obtain for each $n = 2, 3, \cdots$ a sequence $(a_1^{(n)}, \cdots, a_n^{(n)})$ such that $a_1^{(n)} = a$, $a_n^{(n)} = b$ and $a_1^{(n)} < a_3^{(n)} < \cdots < a_n^{(n)}$, and such that if we define $S^{(n)} \equiv [a_1^{(n)}, \cdots, a_n^{(n)}]$, then $S^{(n)} \subset S^{(n+1)}$ for $n = 2, 3, \cdots$. The set-theoretical sum of all the sets $S^{(n)}$ is clearly infinite and linearly ordered, contrary to Lemma 5.1. This completes the proof.

LEMMA 5.3: *If $a \leqq b$, there exists a unique integer $n(a, b) = 1, 2, \cdots$ such that the length m of every principal increasing chain $(b_1, \cdots, b_m)$ with $b_1 = a$, $b_m = b$ is equal to $n(a, b)$.*

PROOF: By Lemma 5.2 there exists an integer $n = 1, 2, \cdots$ and a principal increasing chain $(a_1, \cdots, a_n)$ such that $a_1 = a$, $a_n = b$. Define $n(a, b) \equiv n$. Clearly $n(a, b)$ then has the desired property by Theorem 4.1. To prove the uniqueness, let $n'(a, b)$ along with $n(a, b)$ have the properties in the lemma. Since by Lemma 5.2 there exists $m = 1, 2, \cdots$ and a principal increasing chain $(c_1, \cdots, c_m)$ with $c_1 = a$, $c_m = b$, it is evident that $n(a, b) = m = n'(a, b)$ by Theorem 4.1. This completes the proof.

LEMMA 5.4: *If $a \leqq b \leqq c$, then $n(a, b) + n(b, c) = n(a, c) + 1$.*

PROOF: Select principal increasing chains $(a_1, \cdots, a_m)$ and $(b_1, \cdots, b_n)$ with $a_1 = a$, $a_m = b_1 = b$, $b_n = c$, whence $m = n(a, b)$, $n = n(b, c)$. Then $(a_1, a_2, \cdots, a_m = b_1, b_2, \cdots, b_n)$ is a principal increasing chain of length $n(a, c)$. But its length is clearly $n(a, b) + n(b, c) - 1$. Hence the conclusion follows.

DEFINITION 5.1: *For every $a \in L$ define $d(a) \equiv n(0, a) - 2$. Define also $N \equiv d(1) = n(0, 1) - 2$.*

LEMMA 5.5:

(a) *If $a < b$ then $d(a) < d(b)$;*

(b) *If $a \prec b$ then $d(a) + 1 = d(b)$.*

PROOF: Since $a < b$, $n(a, b) > 1$. Hence $d(b) = n(0, b) - 2 = n(0, a) + n(a, b) - 3 = d(a) + n(a, b) - 1 > d(a)$. If $a \prec b$, then $n(a, b) = 2$, and $d(b) = d(a) + 1$.

LEMMA 5.6: *If $a \leqq b$, then $d(b) - d(a) = n(a, b) - 1$.*

PROOF: Clearly $d(b) - d(a) = n(0, b) - n(0, a) = n(0, a) + n(a, b) - n(0, a) - 1 = n(a, b) - 1$ by Lemma 5.4.

LEMMA 5.7: *For every $a, b \in L$,*

$$d(a + b) + d(ab) \leqq d(a) + d(b). \tag{5.2}$$

PROOF: Let $(a_1, \cdots, a_n)$, $n = n(ab, b)$, be a principal increasing chain with $a_1 = ab$, $a_n = b$. Hence

$$a = ab + a = a_1 + a \leqq a_2 + a \leqq \cdots \leqq a_n + a = a + b,$$

where by Lemma 4.3 either $a_i + a = a_{i+1} + a$ or $a_i + a \prec a_{i+1} + a$ for each $i = 1, \cdots, n-1$. Thus if we delete from the system $(a_1 + a, \cdots, a_n + a)$ each term $a_i + a$ equal to its predecessor we obtain a principal increasing chain $(b_1, \cdots, b_m)$, $m = n(a, a+b)$, with $b_1 = a$, $b_m = a + b$ whence $n(a, a+b) \leqq n(ab, b)$. Now by Lemma 5.6 $d(a+b) - d(a) = n(a, a+b) - 1$, and $d(b) - d(ab) = n(ab, b) - 1$, whence $d(a+b) - d(a) \leqq d(b) - d(ab)$, and our conclusion follows.

LEMMA 5.8: *If* $(a, b)M$, *then* $d(a+b) + d(ab) = d(a) + d(b)$.

PROOF: Let $(c_1, \cdots, c_n)$, $n = n(ab, a)$, be a principal increasing chain with $c_1 = ab$, $c_n = a$. If $ab = 0$, then $(a, b)\perp$; since $c_i \leqq a$, $(c_i, b)\perp$ by Theorem 1.1, whence $(c_i, b)M$. If $ab \neq 0$, then $ab \leqq c_i b$, whence $c_i b \neq 0$, and $(c_i, b)M$ by Axiom B. Thus $(c_i, b)M$ in both cases, and it follows that $c_i + b \prec c_{i+1} + b$ for $i = 1, \cdots, n$ by Lemma 4.4 (since $ab \leqq c_i b \leqq ab$ and therefore $c_i b = c_{i+1} b = ab$). Hence $(c_1 + b, \cdots, c_n + b)$ is a principal increasing chain with $c_1 + b = b$, $c_n + b = a + b$. This chain therefore has length $n(b, a+b)$, and consequently $n(ab, a) = n(b, a+b)$; by Lemma 5.6 $d(a+b) - d(a) = d(b) - d(ab)$, whence the conclusion follows.

LEMMA 5.9: *If* $d(a+b) + d(ab) = d(a) + d(b)$, *then* $(a, b)M$.

PROOF: Let $c \leqq b$. By Lemma 5.7

$$d((c+a)b) \leqq d(c+a) + d(b) - d(a+b+c).$$

Since $a + b + c = a + b$, and since by the hypothesis $d(a+b) = d(a) + d(b) - d(ab)$, we obtain

$$d((c+a)b) \leqq d(c+a) - d(a) + d(ab) = \xi + d(c+ab),$$

where

$$\xi \equiv d(c+a) + d(ab) - d(ab+c) - d(a).$$

Now $ab \leqq ab + c$, whence $ab = a \cdot ab \leqq a(ab+c)$. Thus by Lemma 5.5(a)

$$d(c+a) + d(ab) \leqq d(c+a) + d((ab+c)a)$$
$$= d((ab+c) + a) + d((ab+c)a) \leqq d(ab+c) + d(a),$$

the final inequality holding by Lemma 5.7. Therefore $\xi \leqq 0$, and it follows that $d((c+a)b) \leqq d(c+ab)$. Now if $(c+a)b > c + ab$, we have by Lemma 5.5 $d((c+a)b) > d(c+ab)$, contrary to the result just obtained. In view of the obvious relation $(c+a)b \geqq c + ab$, it follows that $(c+a)b = c + ab$. This proves that $(a, b)M$.

REMARK: The lemma just proved is a generalization of a well-known result that the existence of a real-valued function $d(a)$, $a \in L$ with the properties $d(a) < d(b)$ for $a < b$, and $d(a+b) + d(ab) = d(a) + d(b)$ for every $a, b \in L$ implies that L is modular.[12] The proof of our lemma depends only on the existence of the function $d(a)$, $a \in L$ with the properties stated in Lemma 5.5 (a) and in Lemma 5.7.

[12] Cf. the proof of V. Glivenko, Amer. Jour. Math. 59 (1937), pp. 943, 944.

THEOREM 5.1: *There exists an integer* $N = -1, 0, 1, 2, \cdots$, *and a function* $d(a)$, $a \in L$ *with the following properties*:

(a) $d(0) = -1, d(1) = N$;

(b) *the range of* $d(a)$ *is the set of integers* $[-1, 0, 1, 2, \cdots, N]$;

(c) $d(a) < d(b)$ *for* $a < b$;

(d) $d(a) + 1 = d(b)$ *for* $a \prec b$;

(e) $d(a + b) + d(ab) \leqq d(a) + d(b)$ for $a, b \in L$;

(f) $d(a + b) + d(ab) = d(a) + d(b)$ *if and only if* $(a, b)M$.

PROOF: The integer N and the function $d(a)$ are to be defined as in Definition 5.1. Property (a) is immediate from this definition; parts (c) and (d) are restatements of Lemma 5.5, while parts (e) and (f) were treated in Lemmas 5.7, 5.8, 5.9. To prove (b) it suffices to observe that the range of $d(a)$ is contained in the set $[-1, 0, 1, 2, \cdots, N]$, and that there exists a principal increasing chain $(a_1, \cdots, a_{N+2})$ with $a_1 = 0$, $a_{N+2} = 1$, whence $d(a_i) = i - 2$ for $i = 1, \cdots, N + 2$.

REMARK: That our results are generalizations of those of Dedekind and G. Birkhoff is readily seen from Theorem 5.1 (f), since if L is modular, then the function $d(a)$ satisfies the strong additivity relation $d(a + b) + d(ab) = d(a) + d(b)$ for every $a, b \in L$. If L is a projective geometry or the lattice of linear subsets of an affine space (cf. the example in Section 3), then the function $d(a)$ is precisely the well known geometric dimension function for such a system, as may be easily shown.

Some problems suggested by the results and methods of this paper are the following: (a) To study the topological properties of a lattice L over which there is defined a real-valued function $d(a)$, $a \in L$ with properties (c) and (e) of Theorem 5.1; a treatment of some topics in this connection by M. F. Smiley and the author was recently given.[13] (b) To characterize those semi-modular lattices which arise from modular lattices in the manner described in the example of Section 3; this problem is as yet unsolved. (c) To find and study reasonable non-modular analogs of the continuous geometries of von Neumann; several attempts have been made by the author to solve this problem, but without satisfactory success.

THE INSTITUTE FOR ADVANCED STUDY.

[13] Cf. L. R. Wilcox and M. F. Smiley, *Metric Lattices*, Annals of Math., 40 (1939), pp. 309–327.